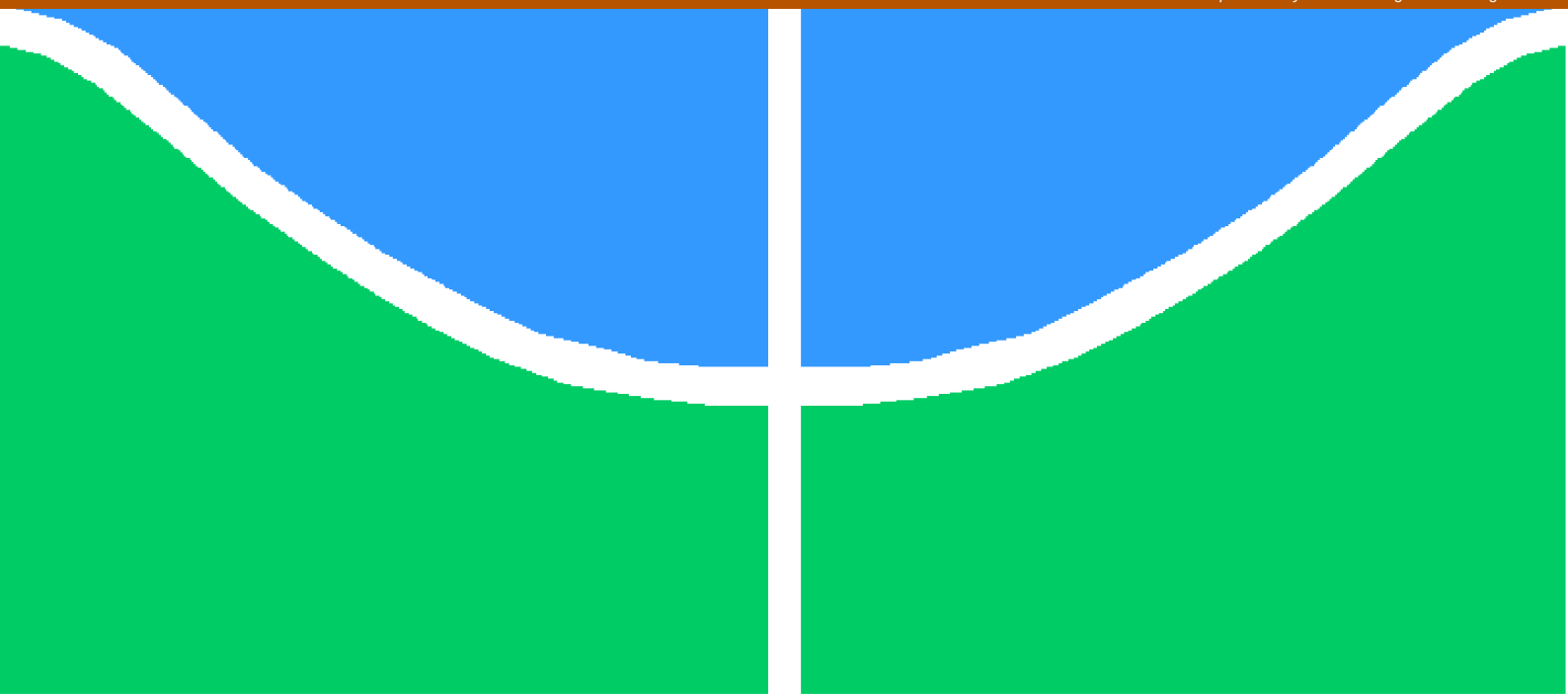


Universidade de Brasília – UnB

Faculdade UnB Gama – FGA

Engenharia de Software

Predizendo epidemias de Dengue, no Distrito Federal, utilizando algoritmos de Regressão

Autor: Pedro Ivo de Andrade, Lucas Vilela Taveira Brilhante

Orientadora: Drª. Carla Silva Rocha Aguiar

Orientador: PhD. Walter Massa Ramalho

Brasília, DF

2018

Pedro Ivo de Andrade, Lucas Vilela Taveira Brilhante

Predizendo epidemias de Dengue, no Distrito Federal, utilizando algoritmos de Regressão

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Universidade de Brasília – UnB

Faculdade UnB Gama – FGA

Orientadora: Dr[a]. Carla Silva Rocha Aguiar

Coorientador: PhD. Walter Massa Ramalho

Brasília, DF

2018

Pedro Ivo de Andrade, Lucas Vilela Taveira Brilhante

Predizendo epidemias de Dengue, no Distrito Federal, utilizando algoritmos de Regressão/ Pedro Ivo de Andrade, Lucas Vilela Taveira Brilhante. – Brasília, DF, 2018-

75 p. : il. (algumas color.) ; 30 cm.

Orientadora: Dr[a]. Carla Silva Rocha Aguiar

Trabalho de Conclusão de Curso – Universidade de Brasília – UnB
Faculdade UnB Gama – FGA , 2018.

1. Machine Learning. 2. Árvores de decisão. I. Dr[a]. Carla Silva Rocha Aguiar. II. Universidade de Brasília. III. Faculdade UnB Gama. IV. Predizendo epidemias de Dengue, no Distrito Federal, utilizando algoritmos de Regressão

CDU 02:141:005.6

Pedro Ivo de Andrade, Lucas Vilela Taveira Brilhante

Predizendo epidemias de Dengue, no Distrito Federal, utilizando algoritmos de Regressão

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Trabalho aprovado. Brasília, DF, 08 de Julho de 2018:

Drª. Carla Silva Rocha Aguiar
Orientadora

PhD. Walter Massa Ramalho
Orientador

Ms. Fabiana Sherine Ganem dos Santos
Orientadora

Ms. Renato Coral Sampaio
Banca Examinadora

Ms. Fábio Macedo Mendes
Banca Examinadora

Brasília, DF
2018

À memória de minha mãe: Sandra, que
sempre se mostrou uma guerreira
ao batalhar pela comida de seus filhos,
e ao mesmo tempo possuía um amor pleno
pela profissão de professora - ensinar

Pedro Ivo

Agradecimentos

Agradeço a minha família, Camila, José Maria, Anna Sofia, Davi "Ceni"e Rodrigo pelo apoio incondicional nessa jornada acadêmica. Agradeço minha madrinha Cléia e o Gustavo por terem me proporcionado um bom colégio no ensino médio, para ingressar na Universidade. Agradeço ao meu padrinho André, pelos conselhos pontuais, ao Rafael, grande amigo dos tempos de UFOP e à República Entrometeu (Ebenezer, Heleno e Oziel) pelos 5 anos de convivência. Agradeço também aos meus orientadores, "Carlove", Walter e Fabiana por depositarem a confiança de realizar este trabalho. Agradeço também a minha dupla de TCC, Lucas Brilhante, que também, é um amigo de confiança.

Pedro Ivo

Agradeço aos meus amigos, ao meu amor Mikhaelle Carvalho, minha mãe Kátia e meu padrasto José Lima que me deram força e apoio durante todo esse tempo.

Lucas Brilhante

"One can't predict the
weather more than a few
days in advance"
Stephen Hawking

Resumo

O dengue é um dos maiores problemas de saúde pública do Brasil. Desde sua origem na África e Ásia, até a descoberta do vírus, foram desenvolvidas muitas pesquisas. Atualmente, muitas abordagens utilizando aprendizado de máquina foram desenvolvidas na Ásia, Brasil e México. Todas elas utilizam bases de dados com registros de Dengue no intuito de observar um comportamento do número de registros de dengue ao longo do tempo. Para essa pesquisa utiliza-se de dados ambientais como precipitação, umidade e temperatura do Instituto Nacional de Meteorologia, além dos registros de dengue provenientes do Sistema de Informação de Agravos de Notificação que serão relacionados pelas datas de seus registros. Utilizando algoritmos de regressão busca-se com esses dados predizer o número de casos de Dengue que ocorrerão no ano de 2014, treinando o algoritmo com dados de 2007 até 2013. Os resultados esperados são um modelo de predição que consiga acertar mais de 80% dos casos de Dengue, além de estabelecer uma função de predição que antecipe o número de casos de dengue e tal modelo seja utilizado por Órgãos responsáveis. Os resultados obtidos foram satisfatórios desde que não use somente dos dados ambientais e do Dengue puramente. É necessário que se realimente o modelo com dados de semanas anteriores para que haja um aperfeiçoamento do treinamento. Além disso observou-se que os algoritmos ADA, GRAD e *Random Forest* foram os que obtiveram melhores resultados.

Palavras-chave: Aprendizado de Máquina, Coleta de Dados, Dengue, Algoritmos de Regressão.

Abstract

Dengue is one of the biggest public health problems in Brazil. Since its origin in Africa and Asia, until the discovery of the virus, many researches have been developed. Currently, many approaches using machine learning have been developed in Asia, Brazil and Mexico. All of them use databases with Dengue records in order to observe a behavior of the number of dengue registries over time. For this research we use environmental data such as precipitation, humidity and temperature of the National Meteorological Institute, in addition to the records of dengue from the Information System of Notifiable Diseases that will be related by the dates of their records. Using regression algorithms, this data is used to predict the number of Dengue cases that will occur in 2014, training the algorithm with data from 2007 to 2013. The expected results are a prediction model that can achieve more than 80 % of Dengue cases, in addition to establishing a prediction function that anticipates the number of dengue cases and such a model is used by responsible organs. The results obtained were satisfactory provided that they did not use only the environmental data and Dengue purely. It is necessary to realign the model with data from previous weeks so that there is an improvement of the training. In addition it was observed that the ADA, GRAD and Random Forest algorithms were the ones that obtained the best results.

Key-words: Machine Learning, Data collecting, Risk Classification, Algorithms, Decision Tree

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ADA	Ada Boosting
ANN	Artificial Neural Networks
API	Application Programming Interface
DT	Decision Tree
GRAD	Gradient Boosting
INMET	Instituto Nacional de Meteorologia
k-NN	k Nearest Neighbors
MFNN	Multi-layer Feed-forward Neural Networks
ML	*Machine Learning*
MLP	Multi-layer Perceptron
RF	Random Forest
RS	Rough Sets
SVM	Support Vector Machine
vs.	Versus

Lista de símbolos

ϵ	Experiência
ρ	Performance
τ	Tarefa

Sumário

1 Introdução

1.1 Motivação

É de conhecimento do meio científico que o Brasil, desde a década de 80, enfrenta muitas dificuldades no controle do Dengue. Hoje, evidencia-se que tal vírus determina um dos maiores problemas de saúde pública, razão pela qual há a necessidade de novas pesquisas, produtos, técnicas, inovações e invenções para eliminar ou erradicar tal doença (TEIXEIRA, 2008).

Um dos fatores que dificultam o controle do vetor do Dengue (o artrópode *Aedes aegypti*) justifica-se no comportamento de Agentes Comunitários de Endemias (ACE), pois muitas vezes suas atividades consistem em intervenções no ambiente privado das famílias, descartando objetos, criticando hábitos, até mesmo culturais, sendo que tais Agentes tem papel fundamental no controle prescritivo, a população tende a ter receio de recebê-los em suas casas (TEIXEIRA, 2008).

Segundo (TEIXEIRA, 2008), é notório que a comunicação, educação e mobilização social, por si só, não são eficazes em produzir mudanças e controlar os problemas de saúde. Para tanto são necessários abordagens muito bem estruturadas, somadas aos hábitos domésticos, como vigilância epidemiológica, combate físico, químico e biológico e sobretudo ações de saneamento básico, além de abordagens transdisciplinares envolvendo diversas áreas do conhecimento.

Desde os anos 2000 até hoje, são desenvolvidos estudos utilizando redes neurais, algoritmos de aprendizado de máquina, até mesmo modelos estatísticos, que tem como intuito contribuir ao combate do Dengue e reverter esse quadro alarmante que a doença paira à nação. Esta pesquisa utiliza de uma combinação de dados de casos de dengue confirmados, além de atrelar à esses casos de dengue variáveis meteorológicas, como temperatura, umidade, precipitação, dentre outras para que possam refinar o modelo.

Foram escolhidos diversos algoritmos de aprendizado de máquina, tanto de classificação, quanto de regressão, que já se provaram eficientes em estudos realizados no Oriente como a pesquisa de (MEHERWAR; MARUF, 2017). Valendo-se da abordagem que utiliza casos de dengue cruzados com as variáveis ambientais, (YUSOF; MUSTAFFA, 2011) também desenvolveu um trabalho semelhante em uma província da Malásia, que mostrou bons resultados e também enfrenta dificuldades na prevenção da doença.

1.2 Justificativa

O Dengue caracteriza, no Brasil, um dos maiores riscos para a saúde pública, há a necessidade de antever surtos de dengue, para que o Governo Federal saiba quando e com que fatores (campanhas de conscientização, combate ao vetor, saneamento básico) devem-se gastar mais recursos econômicos. Discutindo ainda sobre a linha econômica/social, se para o Estado é interessante arcar com os custos do tratamento do Dengue ao invés de preveni-la, não se pode negligenciar a vida daqueles que sucumbem à doença e ao mosquito, pois a parcela de culpa não é somente dos indivíduos infectados.

1.3 Problema de Pesquisa

É possível predizer o número de casos de dengue, no Distrito Federal, de uma semana qualquer, utilizando somente variáveis ambientais e registros de casos de dengue, com algoritmos de aprendizado de máquina?

Para auxiliar a responder essa questão principal, questiona-se outros problemas mais granulares para que se componham na argumentação da questão principal:

1. **Os algoritmos de regressão são adequados para a predição?**
2. **A predição consegue antecipar semanas com muitos casos de dengue?**
3. **Somente as variáveis de precipitações, umidade, temperatura, vento e insolação são suficientes para a predição?**

1.4 Objetivos

1.4.1 Objetivo Geral

Estabelecer um modelo, em que, dadas as variáveis de entrada como, número da semana e dados ambientais, este modelo busque uma relação entre tais variáveis e mostre em formato de gráfico o número de casos de dengue que ocorrerão na semana da entrada.

1.4.2 Objetivos Específicos

1. **Obter dados do site do INMET, para preenchimento das variáveis ambientais;**
2. **Obter dados do Ministério da Saúde com o número de ocorrências de dengue nos últimos anos;**

3. **Tratar dados incompletos, incoerentes e em branco;**
4. **Manipular as datas, para que sejam descritas como o número da semana anual (0-52);**
5. **Excluir *features* que não são relevantes aos algoritmos;**
6. **Testar diversos tipos de algoritmos de regressão e classificação.**

2 Revisão bibliográfica

2.1 Dengue

2.1.1 Etiologia e Histórico

Segundo o *Diccionario de la Lengua Española* da *Real Academia Española*, existem três acepções para definir a palavra dengue. Pelo texto integral,

> **dengue**. (De la onomat. deng, del balanceo.) **1.** Melindre mujeril que consiste en afectar delicadezas, males y, a veces, disgusto de lo que más se quiere o desea. **2.** Esclavina de paño, que llega hasta la mitade de la espalda, se cruza por el pecho, y las puntas se sujetan detrás del talle. Es prenda de mujer. **3.** Pat. Enfermedad febril, epidémica y contagiosa, que se manifiesta por dolores de los miembros y un exantema semejante al de la escarlatina."

No ano de 1891, o médico *Schuchardt*, baseando-se no comportamento dos enfermos e a acepção de meneio, balanço (REZENDE, 2004), ressalta que a doença foi assim denominada,

> "sea por la tiesura que dejan los dolores del dengue, sea porque el dengue que a veces es enfermedad leve, fuese tachado por alguns de mera afectatión"

Antes da classificação formal da doença, ela era determinada por regionalismos, no Brasil. Haviam-se várias denominações como "patuléia", "polca", "urucubaca", "melindre"e "febre eruptiva reumatiforme"(OSANAI, 1984).

O Dengue tem origem silvestre e foi, primordialmente, mantido por primatas inferiores, na Ásia e África (GUBLER, 1998). Nos últimos séculos, a doença evoluiu, e passou a estabelecer uma associação epidemiológica estritamente antropofílica, estando relacionada a pandemias e epidemias em todo o mundo (TEIXEIRA; BARRETO; GERRA, 1999), principalmente pela intensa movimentação de pessoas e cargas entre os portos desde o tempo das grandes navegações (OPAS, 1995).

Em sua dissertação de mestrado (RAMALHO, 2008) afirma o seguinte:

> A dengue é uma doença infecciosa não contagiosa, causada por arbovírus, ou vírus transmitido por artrópodes, e distingue-se quatro sorotipos, o DEN-1, DEN-2, DEN-3 e o DEN-4, antigenicamente diferentes, e cada grupo ainda compostos por subcomplexos sorológicos.

2.1.2 O vetor

O vetor do Dengue é o artrópode *Aedes aegypti*, originário da região mediterrânea, principalmente do Egito. Seu ciclo biológico é caracterizado nas condições climáticas favoráveis da região tropical, onde vive em média de 8 a 12 dias, compreendendo as fases de ovo, os 4 estágios larvais, pupa e adulto conforme explicam (RAMALHO, 2008) e (SANTOS, 2003).

Seu desenvolvimento embrionário completa-se entre 48 e 72 horas após a oviposição, a eclosão dos ovos liberam larvas que evoluem em pupas entre 2 a 5 dias, posteriormente se transformam em adultos entre 2 a 3 dias, e por fim, em sua fase adulta, podem viver em média 45 dias (SANTOS, 2003).

Devida à sua grande capacidade de adaptação ao meio urbano, os criadouros do vetor caracterizam por recipientes preenchidos com água de chuvas, como, calhas, lajes, pneus, latas, pedaços de plástico, ou ainda recipientes para armazenar água em uso doméstico como as cisternas, caixas d'água, vasos de plantas, e tanques mal tampados (TEIXEIRA; BARRETO; GERRA, 1999). Sabe-se ainda que as fêmeas podem utilizar o meio natural como criadouros, onde buracos de árvore, bromélias, interior de bambus podem ser recipientes favoráveis ao seu desenvolvimento (FORATTINI, 1994).

Sabe-se ainda, que as fêmeas restringem seus hábitos hematófagos aos horários diurnos, cujos picos de atividade são caracterizadas ao amanhecer ou no anoitecer (FORATTINI, 1994). O mosquito tem temperamento arisco e consegue esquivar-se da vítima quando está em perigo. Permanecendo insaciado, pode procurar outras vítimas e fazer múltiplas ingestões de sangue, e tal ato é que amplia as chances de infectar-se ou transmitir o vírus, conforme (FORATTINI, 1994).

2.1.3 Biometeorologia e Mudanças Climáticas

Segundo o dicionário Aurélio a Biometeorologia tem como objetivo estudar as relações ou influências direta ou indiretas da atmosfera nos organismos vivos. Neste contexto, busca-se entender as relações entre o *Aedes aegypti*, o vírus e como podem ser influenciados pela temperatura, umidade, chuvas e outros fatores atmosféricos (SCHREIBER, 2001).

No período de oviposição é entendido que o mosquito tem altas taxas desse índice quando a temperatura do ambiente está entre 20°C e 30°C, para temperaturas abaixo de 18°C ou acima de 34°C têm-se uma diminuição da fecundidade das fêmeas, em que as mesmas guardam suas energias apenas para a sobrevivência (SCHREIBER, 2001).

Define-se como período de incubação extrínseco (PIE) o intervalo de tempo em que o vírus se desenvolve dentro do mosquito, e possui relação não linear com altas temperaturas (PATZ *et al.*, 1998). O aumento da temperatura induz numa diminuição do

PIE, pois enquanto o DEN-2 a 30ºC necessita de 12 dias para se desenvolver, a 32-35ºC requer apenas 7 dias. Além disso, uma vez que o vetor se torna infectante ele permanece em tal estado pelo resto de sua vida (PATZ *et al.*, 1998).

Define-se como período de incubação intrínseco (PII) o intervalo de tempo em que o vírus se desenvolve no ser humano (PATZ *et al.*, 1998). O intervalo pode ser de 3 a 15 dias com variações entre 5 e 6 dias, e a viremia inicia-se no dia anterior do aparecimento da febre, e permanece até o sexto dia (PATZ *et al.*, 1998).

Tratando-se das diversas variáveis que influenciam na facilidade do desenvolvimento do mosquito e consequentemente do vírus, temos os fatores ambientais, temperatura e umidade. Estudos realizados por (WU *et al.*, 2007), em temperaturas altas (28ºC) e baixa umidade (50-55%) o ambiente mostra-se mais favorável ao mosquito, pois, os vetores aumentam sua busca por alimentos, comparando-se com temperaturas mais baixas (25ºC) e mais alta umidade (85-90%).

2.2 Aprendizado de Máquina

Muitas vezes em contextos do mundo real problemas são resolvidos baseando-se na capacidade de decisão de uma pessoa. Essa capacidade vem do conhecimento e da experiência em determinada área. Como no exemplo citado por (BRINK; RICHARDS; FETHEROLF, 2016), um gerente bancário tem a decisão de conceder um empréstimo ou não para uma pessoa, baseado na ficha de informações e histórico dela. Com estas informações o gerente bancário tenta prever, apoiado em sua experiência, se aquela pessoa irá ou não honrar seus compromissos de pagar a dívida. A identificação desses padrões, para prever o comportamento do cliente neste contexto, está limitado a capacidade da mente humana. Uma abordagem mais eficaz seria a utilização de aprendizado de máquina.

Ao final dos anos 50, no ano de 1959, especificamente, (SAMUEL, 1959) define *Machine Learning* (Aprendizado de Máquina) como sendo uma área de estudo que propicia ao computador a habilidade de aprender sem ser explicitamente programado para tal fim. (SAMUEL, 1959) em seus estudos realizava a implementação de uma inteligência artificial para um jogo de damas, em que, buscava a partir de uma árvore de decisão executar a melhor jogada com base na posição de cada peça.

Alguns anos depois, na década de 90, (MITCHELL, 1997) define o Aprendizado de Máquina da seguinte maneira:

Definição 1. *Um programa de computador é dito para aprender com a experiência ϵ em relação a alguma tarefa τ e alguma medida de desempenho ρ, se seu desempenho em τ, conforme medido por ρ, melhora com a experiência ϵ.*

Pode-se exemplificar tal definição com a utilização de um *software* anti-spam em

uma caixa de *e-mails*. A ação do *software* classificar o *e-mail* em *spam* ou não-*spam* representa a tarefa τ; o *software* observar um conjunto de dados rotulados como *spam* ou não-*spam* é dado como a experiência ϵ; e o número ou porcentagem de *e-mails* corretamente classificados como *spam*/não-*spam* é a performance ρ. Segundo (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012), este tipo de aprendizagem é tipicamente utilizada em algoritmos de classificação, regressão e problemas de *ranking*.

Em sua visão, (MICHIE; SPIEGELHALTER; TAYLOR, 1994), define *Machine Learning* também como:

Definição 2. *O ato de ensinar a máquina a fazer decisões baseadas em dados apresentados. Diversas implementações apresentam aprendizado a partir de operações lógicas ou binárias.*

O autor (BRINK; RICHARDS; FETHEROLF, 2016) complementa a definição anterior da seguinte maneira:

Definição 3. *O papel do Machine Learning é identificar padrões e a partir deles predizer qual será o próximo resultado dada uma nova entrada.*

A aprendizagem de máquina se divide basicamente em duas grandes áreas, são elas, **Aprendizado Supervisionado** e **Aprendizado Não-Supervisionado**. Há autores como (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012) que, além das duas áreas previamente citadas ainda acrescenta mais 3 subdivisões como, **Aprendizagem Semi-Supervisionada**, **Inferência Transdutiva** e **Aprendizagem *Online***. Há controvérsias sobre tal divisão, pois na verdade elas são ou uma mesclagem de aprendizagem supervisionada com não-supervisionada, ou uma derivação delas.

Quando falando de Aprendizagem de máquina, alguns termos serão se suma importância e serão recorrentes neste capítulo de revisão teórica e inerentes ao aprendizado de máquina e suas formas de realização(?). Uma *feature*, segundo o glossário de termos de (KOHAVI; PROVOST, 1998) definimos,

Definição 4. ***Feature(Característica)*** **-** *Representa a quantidade que descreve uma instância; uma especificação de um atributo e seu valor. Uma cor pode ser considerada um atributo, exemplifica-se uma tal termo como "A cor é azul".*

Uma *feature* possui um domínio definido pelo seu tipo, que denota quais valores podem estar associados ao atributo. Os domínios se subdividem em dois tipos, Categórico e Contínuo.

Domínios categóricos são um conjunto finito de valores discretos, pode-se citar o tipo nominal que caracteriza valores sem ordem, como nomes e cores, e o tipo ordinal

que denota ordem, como baixo, médio ou alto. Domínios contínuos são descritos como um subconjunto dos números reais, onde há uma diferença considerável entre os valores possíveis de serem atribuídos.

Ao longo dos anos um problema recorrente visto na utilização de *Machine Learning* é o uso indiscriminado dos dados crus. Isto resulta em modelos viciados ou pouco acurados. Este problema se torna tão recorrente porque a coleta de dados não necessariamente foi visando a utilização de uma algoritmo de aprendizagem, então antes de começar o aprendizado é importante a prática de *Feature Selection*, engenharia de característica e limpeza de dados inválidos. (HACKELING, 2014) classifica o conjunto dessas práticas como o pré-processamento, uma etapa essencial na modelagem do problema de *Machine Learning.*

Para (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012) define-se uma *label* e exemplo de treinamento da seguinte forma,

Definição 5. ***Label -*** *Valores ou categorias atreladas à um* ***exemplo de treinamento****. São os valores relacionados a um conjunto de características. O valor ou categoria a ser encontrado pelo modelo de ML. A label é utilizada para o treinamento quando utilizamos aprendizado supervisionado.*

Por exemplo: para as características no contexto problema do banco apresentado: - Valor do empréstimo 5000, Homem, 35 anos, volume bancário acima de 200k/ano, Nome 0 vezes no Serasa - O rótulo dessa entrada seria: Aprovado. Porém se as características fossem: Valor do empréstimo 20000, Homem, 29 anos, volume bancário acima de 100k/ano, Nome 2 vezes no Serasa - O rótulo dessa entrada seria: Rejeitado. Se apresentado dados dessa forma, o rótulo teria dois possíveis valores, dois resultados possíveis: Aprovado e Rejeitado, um exemplo de categorização. Para (HACKELING, 2014) temos as definições de dados de entrada em algoritmos de *Machine Learning*:

Definição 6. ***Base de Treinamento:*** *São itens ou instâncias de dados utilizadas para o aprendizado de máquina.*

Definição 7. ***Base de Teste:*** *São itens ou instâncias de dados utilizadas para a avaliação da precisão do aprendizado de máquina.*

Definição 8. ***Base de Validação:*** *São itens ou instâncias de dados utilizadas para validar o medelo na prática. Normalmente os ultimos dados são selecionados, simulando a aplicação no mundo real.*

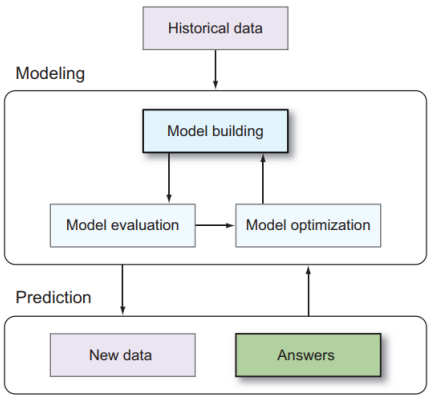
2.2.1 Processo de *Machine Learning*

Antes de começar a analisar o problema e entender a modelagem do aprendizado, é necessário saber quais passos são imprescindíveis quando trata-se de Aprendizado de

Máquina.

Durante os últimos 30 anos tivemos um crescimento incrível na aplicação de machine learning. Esse crescimento rendeu um processo consolidado por qual toda aplicação deve passar, (BRINK; RICHARDS; FETHEROLF, 2016). Podemos importar uma imagem do livro "Real World Machine Learning"para entender melhor o processo.

Figura 1 – Fluxograma básico de funcionamento de *Machine Learning*



Fonte: (BRINK; RICHARDS; FETHEROLF, 2016)

O fluxo de trabalho durante o processo de resolução de problemas de ML consiste nas seguintes etapas: Modelagem e Predição.(BRINK; RICHARDS; FETHEROLF, 2016)

Durante a etapa de modelagem tem-se os subprocessos de *Construção do Modelo*, *Avaliação do Modelo* e *Optmização de Modelo*, a Tabela 1 descreve as etapas propostas.

Tabela 1 – Etapas do processo de modelagem de Machine Learning

ETAPA	DESCRIÇÃO
Construção do Modelo	É aqui onde é analisado o contexto, os dados, se já existirem, e o problema. Para uma análise bem orientada também é necessário que nesta atividade seja feita uma pesquisa de problemas parecidos e algoritmos que atendam as características do contexto. Aqui também é onde deve haver o tratamento dos dados para auxiliar a utilização dos algoritmos. Dados vazios, não parametrizados, ou não normalizados podem atrapalhar no funcionamento dos algoritmos. Este subprocesso é onde este trabalho será focado.
Avaliação do Modelo	Este subprocesso é onde há a execução do modelo e a aferência sobre os resultados. No caso de algoritmos supervisionados, podemos identificar a taxa de acerto do algoritmo sobre a base de treino e a base de teste, caso exista dados já coletados.
Optimização do Modelo	Após o recebimento de novos dados e de ter sido feita a predição, devemos identificar se o modelo esta tendo uma taxa de acerto dentro do esperado. Com estas informações podemos tomar decisões quanto a melhora ou mudança do modelo que está sendo utilizado.

Já durante a etapa de *Predição* podemos observar onde o algoritmo é utilizado e testado. Novos dados são colocados no modelo e novas respostas são obtidas. Estas respostas servirão de entrada para o subprocesso de *Optimização de Modelo*.

2.2.2 Aprendizado Não-Supervisionado

Definição 9. *Algoritmos não-supervisionados não possuem rótulos, ou seja, não existe uma resposta pronta para um dado conjunto de características. O algoritmo não-supervisionado funciona encontrando padrões a partir das características, identificando o que normalmente ocorre e o que não ocorre no conjunto de dados, (HU; HAO, 2012).*

Segundo os autores (KOHAVI; PROVOST, 1998), o aprendizado não-supervisionado é uma técnica que busca agrupar instâncias sem haver entre elas atributos pré-especificados dependentes.

De acordo com (LOURIDAS; EBERT, 2016), no aprendizado não-supervisionado tem-se apenas dados e não suas soluções relacionadas, isto quer dizer que a própria má-

quina deve encontrar soluções para uma predição, os autores exemplificam de tal forma o aprendizado

> "É como dar a um estudante um conjunto de padrões e pedir-lhe para que descubra os motivos subjacentes que geraram tal padrão."

Na visão de (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012), no aprendizado não-supervisionado, geralmente, o aprendiz (máquina) recebe um conjunto de dados sem *labels* e faz a predição para tais dados. Como os dados sem *labels* são difíceis de serem manipulados, tende-se a encontrar problemas de performance na máquina, pode-se utilizar da técnica de *clustering* e redução dimensional para solucionar tal empecilho. Em diversos contextos que não é possível rotular ou coletar os dados de maneira completa, a abordagem não-supervisionada se torna necessária.

2.2.3 Aprendizado Supervisionado

Definição 10. *O aprendizado supervisionado é uma estratégia de aprendizagem caracterizada pela presença de um conjunto de treinamento que possui dados e a saída correta da tarefa relacionada a um dado.*

Conforme define (LOURIDAS; EBERT, 2016). Pode-se exemplificar tal situação da seguinte maneira:

> "Dê a um estudante um conjunto de problemas e suas soluções, então diga a esse aluno que descubra como resolver outros problemas que ele terá que lidar no futuro, com base no conjunto de dados que já lhe foi apresentado."

Tal tipo de aprendizado, segundo (LOURIDAS; EBERT, 2016), pode ser caracterizado por **algoritmos de classificação**, em que ele obtêm como entrada um conjunto de dados, e estes dados possuem classes, então a função deste algoritmo é aprender a classificar novos dados. Outro modo de utilizar o aprendizado supervisionado é através de **algoritmos de regressão**, que visam prever o valor de um atributo de uma entidade com base em análises estatísticas.

Conforme a linha de estudo de (ANTHONY; BARTLETT, 2009), em geral, um sistema de aprendizado supervisionado deve prever as *labels* dos padrões. Durante o treinamento da máquina, ela recebe partes de informações sobre a relação padrão *vs.* labels, na forma de - número de padrões corretamente rotulados (*labeled*). Um exemplo deste tipo de treinamento encontra-se em uma aplicação de reconhecimento de faces,

"em que um sistema recebe um número de imagens e, tal sistema rotula as imagens como um usuário legítimo, ou um impostor."

Apesar de parecer mais intuitivo, o algoritmo supervisionado pode apresentar problemas que a princípio podem não parecer problemas. O mais lógico é o *Underfitting*, que é definido por:

Definição 11. *o algoritmo que foi treinado sem o número necessário de informações para realmente ser um modelo representativo daquele contexto.*

Já o *Overfitting* pode não ser tão aparente, tal problema é concebido pela seguinte proposição

Definição 12. *Ocorre quando o modelo fica viciado na base de treino, apresentando resultado de acerto próximos a 100%, mas quando afrontado com a base de teste falha completamente. Este problema exibe a grande necessidade de conter bases diferentes de treino e de teste.*

Todos estes conceitos foram exemplificados na visão do autor, (HACKELING, 2014).

2.2.4 Predição e Inferência

Quando projetando um modelo de ML, devemos ter um objetivo em mente. O objetivo pode ser Preditivo ou de Inferência.

Ao se utilizar um modelo de ML, em geral, os projetistas possuem uma finalidade, ela pode ser de interesse preditivo ou deducente. O primeiro tem como intuito prever algo, o segundo tem como objetivo inferir uma análise sobre os dados processados. As definições abaixo esclarecem de forma mais conceitual tais conceitos. (BRINK; RICHARDS; FETHEROLF, 2016)

Definição 13. ***Predição:*** *Tem o intuito de prever algum resultado futuro baseado nas experiências passadas. Bons exemplos disso são: Predição do preço da bolsa de valores ou identificar se um consumidor está mais incitado a comprar itens de uma certa categoria.*

Definição 14. ***Inferência:*** *A partir de dados obtidos na predição é possível inferir algo sobre aquele contexto. É neste momento que o ML atinge o seu grande objetivo de ver o que os olhos humanos não conseguem observar. Padrões escondidos nos dados podem dar informações importantes para tomar decisões. Um bom exemplo disso trata-se de uma predição epidemiológica de alguma doença, uma inferência sobre tal contexto poderia prevenir a morte de milhares de pessoas.*

2.2.5 Pré-Processamento

Ao receber dados do mundo real, muitas perguntas a respeito da natureza dos dados são feitas, para assim, determinar o melhor algoritmo de aprendizagem de máquina. Porém, existe um tópico a cerca dos dados igualmente importante: O preparo dos dados. Quando se adquire dados de um determinado contexto, raramente este virá formatado e preparado para a utilização de um método de aprendizado de máquina, ainda mais considerando que os vários métodos existentes, trabalham melhor com diferentes formatações dos dados. (BRINK; RICHARDS; FETHEROLF, 2016)

Dentre essas modificações necessárias temos dados categorizados e numérico, dados inválidos e em branco, normalização e engenharia de característica.

2.2.5.1 Dados Categorizados

Sabemos que um dado é categórico quando os valores podem ser agrupados e a ordem dos valores não tem significado. Porém, quando utilizamos dados categóricos com valores numéricos, podemos estar adicionando peso a certas categorias, dessa forma possivelmente viciando o modelo a dar mais peso para alguns valores. Por exemplo: Se tivermos utilizando os meses do ano, de janeiro a dezembro, como valores de 1 a 12 respectivamente, estaremos possivelmente dando mais peso para dezembro, e o menor peso para janeiro (BRINK; RICHARDS; FETHEROLF, 2016). Para lidar com esse tipo de problema é necessário dividir em valores binários. Por exemplo Tabela 2

Tabela 2 – Gêneros

Sexo
Masculino
Feminino

A *feature* sexo pode ter dois valores: “Feminino” e “Masculino”, ou 0 e 1. Para impedir o problema citado acima podemos converter essa *feature* em Tabela 3

Tabela 3 – Gênero Categorizado

Masculino	**Feminino**
1	0
0	1

De acordo com (BRINK; RICHARDS; FETHEROLF, 2016), outros vários métodos de tratamento que possa ser necessário com o dados antes de colocá-los em um modelo de aprendizagem. Alguns deles podem ser evitados se tratados logo na coleta de dados, como evitando formulários que possuem texto extenso.

Tendo em vista estes conceitos, e lembrando que o foco deste trabalho é no subprocesso de *Construção do Modelo*, pode-se finalmente começar a analisar o problema, os dados e os algoritmos.

2.3 Algoritmos

Tendo em vista que o objetivo deste trabalho foca na utilização dos algoritmos e não no aprofundamento dos mesmos, a seguir serão brevemente descritos algoritmos escolhidos. Vale a pena ressaltar que a escolha dos algoritmos foi visando a criação de modelos de naturezas diferentes, mas que estivessem relacionados com o problema de pesquisa: Predição de surtos de dengue.

2.3.1 *DT (Decision Tree)*

Pode-se definir uma árvore de decisão, conforme diz (MITCHELL, 1997), como um método para aproximar a natureza dos dados - *features* - em funções, onde a função de aprendizagem é representada por uma árvore de decisão. Tais árvores aprendidas podem ser representadas - a nível de código fonte - como conjuntos de estruturas condicionais *"se-então"* para melhorar a leitura e entendimento humano, de acordo com (MITCHELL, 1997).

Podemos exemplificar uma árvore de decisão em que a máquina deve decidir com base nas variáveis do tempo (ensolarado, nublado ou chuvoso), se pode ou não ocorrer uma partida de tênis. Além das variáveis de tempo, têm-se outras variáveis que podem ser levadas em conta dependendo da condição climática local, como umidade (alta ou normal) e o vento (forte ou fraco). Veja a figura 2.

Da mesma forma funciona a *Decision Tree Regressor*, veja a figura .

Figura 2 – Exemplo de árvore de decisão, sobre condições para realização de um jogo de tênis.

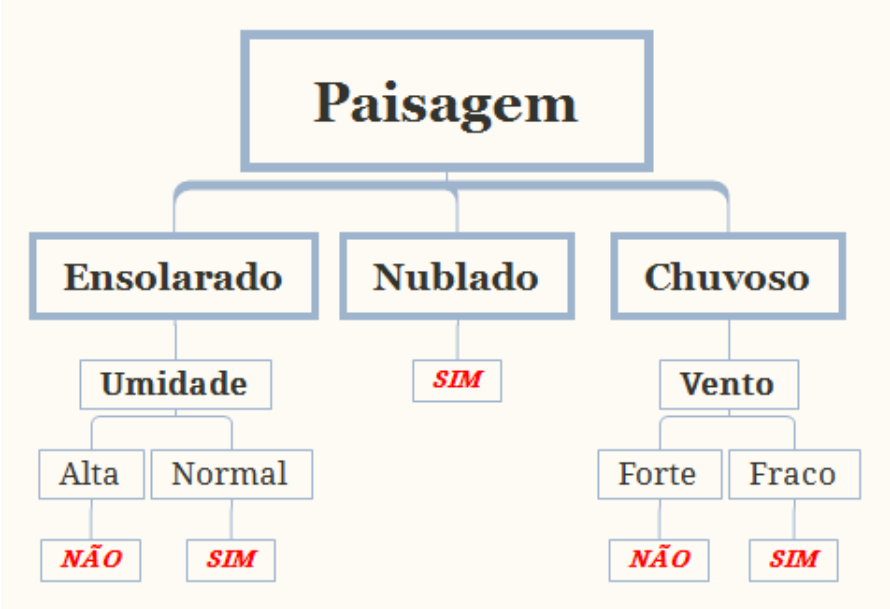
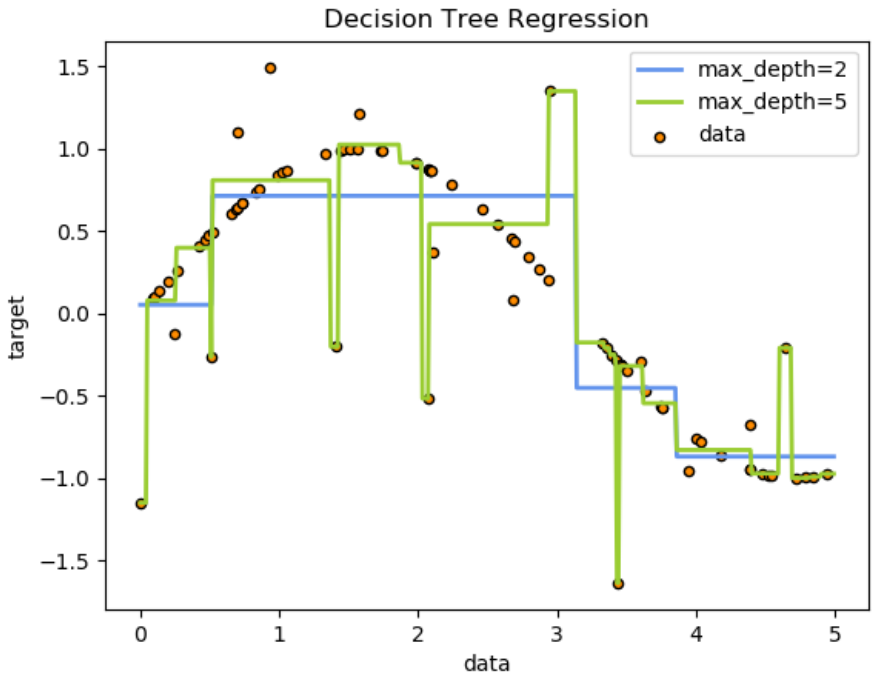


Figura 3 – Funcionamento do *Decision Tree Regressor*

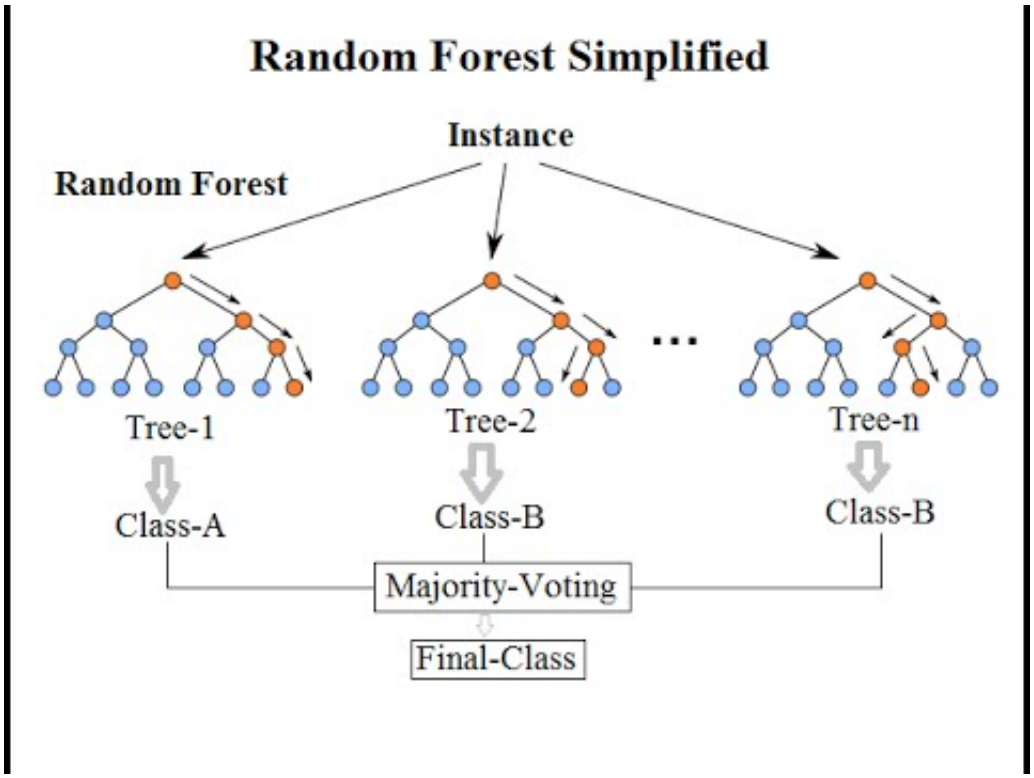


2.3.2 *RF (Random Forest)*

O autor (BREIMAN, 2001) descreve *Random Forest* como uma evolução das *decisions trees*, onde várias árvores são formadas para criar um modelo com maior precisão. Isto é feito a partir da separação dos Dados em conjuntos de dados menores e aleatórios. Cada árvore é construída a partir de um pedaço aleatório dos dados. Quando um novo

dado chega, a predição é feita por todas as Árvores e ao fim é feita uma votação por maioria, ou seja, a categoria com mais votos ganha e o resultado é dado. Podemos entender melhor observando a Figura .

Figura 4 – Simplificação do modelo *Random Forest*



De acordo com (BREIMAN, 2001) as RF (*Random Forest*) corrigem a maior parte dos problemas de *Overfitting* que as Árvores de decisão apresentam. Tudo depende do quanto as DTs contidas dentro da *Random Forest*. Isto é, o quanto elas representam os dados.

As *Random Forests*, assim como as *Decision Trees* são ideais para o problema de classificação hospitalar, pois além de aceitar uma dimensionalidade bem alta, não necessitam da normalização dos dados e também não há a necessidade de um número de dados muito grande para começar-se a apresentação de resultados.

2.3.3 *Adaptative boost*

Adaptative Boost ou *AdaBoost*, é uma técnica para melhorar a performance de algoritmos de *machine learning*. Ele funciona treinando o modelo com a base de dados original e então treina com mais cópias dos dados originais mas modificando os pesos que cada pedaço do dado tem baseado nos erros cometidos pelo modelo já treinado, dessa forma sendo capaz de predizer os casos mais frequentemente errados. O *AdaBoostClassifier* e o *AdaBoostRegressor* presente no *scikit-learn* usa como base para predição o algoritmo *Decision Tree*. (FREUND; SCHAPIRE, 1997).

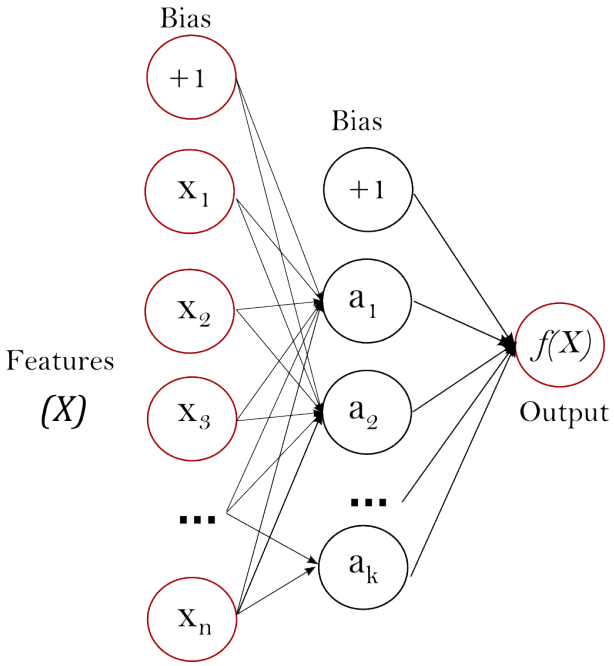
2.3.4 *Gradient boost*

De acordo com (FRIEDMAN, 2001), *Gradient Boosting* pode ser resumida em um algoritmo de otimização que se adéqua a função diferenciável de perda. O GradientBoostingRegressor e GradientBoostingClassifier também usam a decision tree como algoritmo padrão.

2.3.5 *Multi-layer Perceptron* (MLP)

MLP é uma rede neural com múltiplas camadas de neurônios que se utiliza do recurso de *backpropagation*. Um neurônio é composto basicamente de três estruturas: conexões de entrada, combinador linear e função de ativação. O algoritmo de backpropagation, de forma simplista percorre a rede de uma forma específica com o objetivo de ajustar o erro e os pesos das sinapses, que é a ligação entre os neurônios. (RUMELHART; HINTON; WILLIAMS, 1986)

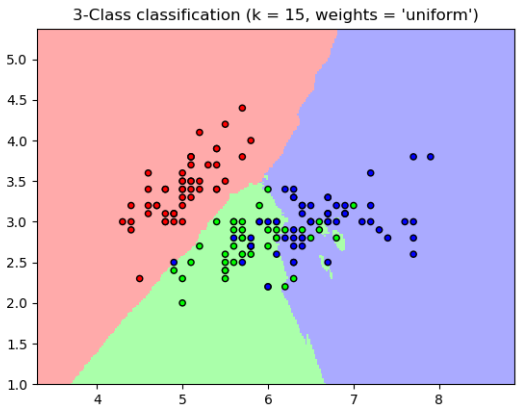
Figura 5 – Funcionamento da MLP



2.3.6 *k nearest neighbors*

O algoritmo k-NN se baseia na interpolação dos k vizinhos mais próximos, que são definidos pelas *features*, quanto mais features, mais espaços dimensionais e mais complexo o funcionamento do algoritmo. Para a regressão normalmente é utilizado a distância euclidiana ou a distância de Mahalanobis para encontrar os vizinhos mais próximos. (PETERSON, 2009).

Figura 6 – Visualização do modelo de classificação *k-NN*



2.4 Trabalhos relacionados

O Dengue, possui vários estudos, conforme apresentado anteriormente, que datam do início do século XVIII, e foram aprimorados ao longo dos anos. Alguns estudos recentes mostram a busca de pesquisadores por um modelo que descreva o comportamento do vírus, qual será seu impacto na saúde pública, e quais fatores podem influenciar em tais aspectos. Os aspectos por detrás de tais modelagens podem ser, geográficos, estatísticos, computacionais e matemáticos, e podem auxiliar Instituições responsáveis a tratar o vírus, controlar sua disseminação, além do próprio avanço da pesquisa sobre o assunto.

Desenvolveu-se por (MEHERWAR; MARUF, 2017), uma pesquisa sobre os algoritmos de Machine Learning para diagnóstico de diversas doenças, dentre elas, a dengue. No estudo utilizou-se uma base de dados do Departamento de Saúde Pública do Estado de Selangor (Malásia), e realizou-se um estudo comparativo da precisão e acurácia de algoritmos de classificação como (*Decision Trees*, ANN, SVM, RS e MFNN) para obter melhores resultados quando trata-se apenas de registros de dengue.

Outra modelagem, diz respeito ao estudo produzido por (MACHADO-MACHADO, 2012), em que busca-se mapear empiricamente a adequação de dengue no México utilizando a distribuição de espécies para elaboração de seu modelo. O estudo também utiliza de fatores climáticos e sócio-econômicos para melhores resultados e estabelece relações entre tais variáveis.

Em uma revisão bibliográfica, realizada por (FATHIMA; MANIMEGALAI; HUNDLEWALE, 2011), estabelece-se o cenário de pesquisas até o momento da publicação, em que são elencados as principais aplicações de ML no contexto de doenças de dengue. O estudo exibe uma série de algoritmos vastamente utilizados nas pesquisas sobre dengue (DTs, SVM, Algoritmos Genéticos, *Evolutionary Programming*, Redes Neurais, *Fuzzy Sets*, e RS), e além disso, estabelece um processo de como são feitas as metodologias, desde a obtenção de dados até a validação dos modelos.

Os autores, (YUSOF; MUSTAFFA, 2011), em um de seus trabalhos buscavam predizer surtos de Dengue, em cidades da Malásia (Sepang, Hulu Selangor, Hulu Langat, Klang, Kuala Selangor), utilizando dados de registros de dengue e volume de chuva da época, para aplicar tais variáveis ao algoritmo de uma (*Least-Squares Support Vector Machine*) LS-SVM, para predizer os anos de 2005 e 2006.

3 Modelagem do Problema

3.1 Os dados

Os dados utilizados são provenientes do INMET (para as variáveis ambientais) e SINAN (Sistema de Informação de Agravos de Notificação) para os registros de atendimentos de dengue no Distrito Federal. Em ambos os casos os dados estão mal formatados e necessitam de alguns tratamentos e padronizações para que sejam utilizados no modelo.

Os dados de registros de Dengue do ano de 2007 a 2014 estão disponíveis em arquivos no formato "csv", e possuem em média 800 registros (anos não epidêmicos) e 10000 registros (anos epidêmicos), em que os registros contabilizam, casos confirmados, não confirmados e negativas. Os campos da tabela descrevem data do primeiro sintoma, data da identificação do sorotipo, município de atendimento, nome do paciente, sexo, endereço do paciente, endereço da unidade de atendimento, resultado do exame, data do exame, dentre outros que estavam em branco. A tabela ?? abaixo exemplifica como é a tabela de dados original.

ID_MUNICIPIO	ID_UNIDADE	DT_SIN_PRIN	SEM_PRI	FONETICA	SOUDEX	CS_SEXO	SG_UF	ID_MIN_RES	ID_DISTRITO	ID_BAIRRO	RA	NUM_BAIRRO
00000001	10320123	20070622	20700	FULANO	E2123123	M	53	530010	27	1	BRASILIA	UMS ASA SUL

Para os dados meteorológicos, obtidos no site do INMET, são disponibilizados em páginas HTML, e os pesquisadores necessitam fazer uma singela seleção dos dados na página e salvá-los em arquivos no formato "csv", para que possam ser manipulados. Os dados contemplam temperatura, umidade, precipitação, evaporação, insolação e velocidade do vento. Também foram utilizados os dados referentes aos anos de 2007 a 2014, para manter a combinação com os registros do Dengue.

Week	Cases	Precipitacao_mean	TempMax_mean	TempMin_mean	Insolação_mean	Evaporação Pinche_mean	Temp Comp Média_mean	Umidade Relativa Média_mean	Velocidade Tempo Media_mean
10	7	71	21.5	17.1	0	3	18.92	92.75	0

3.2 O modelo

3.2.1 Categoria do Modelo

Conforme já foram apresentados os dados que serão utilizados na modelagem, nota-se que são dados rotulados, e as *features* são as colunas da base de dados dos pesquisadores (que mescla o número de casos em uma determinada semana anual juntamente aos seus dados ambientais), conforme a figura ??, sendo assim caracteriza um modelo de ML de caráter **supervisionado**.

Week	Cases	Precipitação	TempMax	TempMin	Insolação	Evaporação Pinche	Temp Comp Média	Umidade Relativa Média	Velocidade Tempo Media
10	7	71	21.5	17.1	0	3	18.92	92.75	0

Em relação a predição dos casos, os pesquisadores buscam responder também, qual é a abordagem mais adequada, **classificação** tratando os resultados de predição como intervalos (0 a 50 casos, 50 a 100 casos, 100 a 200 casos e 200 a 10000 casos), ou a **regressão** que se importa com o número de casos estabelecendo uma espécie de função matemática.

3.2.2 Algoritmos Selecionados

Após a definição básica do funcionamento dos algoritmos no tópico 2.3, deste documento, estabelece-se a justificativa de sua utilização nesta seção. Primeiramente, de acordo com o método de pesquisa, Desenvolvimento Orientado à hipótese, temos a possibilidade de definir uma hipótese e testá-la, por exemplo, "O algoritmo *Random Forest* é melhor para predizer os casos de dengue ao invés do MLP?". Deste modo selecionou-se seis algoritmos, abrangendo classificadores e regressão. Ambos foram selecionados devido à sua utilização em outras pesquisas conforme descrito no tópico 2.4, e comprovados que eram eficientes para abordagens de contextos de Dengue. Dessa maneira, os algoritmos a serem testados na modelagem foram:

- Decision Tree;
- Random Forest;
- k-Nearest Neighbors;
- Multi-Layer Perceptron;
- ADA Regressor;
- GRAD Regressor.

3.3 A solução

Para auxiliar os Órgãos Governamentais responsáveis pela tomada de decisão sobre o controle e erradicação do vetor do Dengue, propõe-se o estabelecimento de um modelo preditivo, utilizando algoritmos de ML, possuindo como entrada variáveis ambientais e dados sobre registros de dengue, seja capaz de prever o número de casos de dengue que irá ocorrer nas semanas seguintes.

4 Metodologia

4.1 Método de Pesquisa

4.1.1 Natureza da Pesquisa

> Essencialmente, a pesquisa experimental consiste em determinar um objeto de estudo, selecionar variáveis que seriam capazes de influenciá-lo, definir as formas de controle e de observação dos efeitos que a variável produz no objeto, (GIL, 2010).

Para garantir a pesquisa experimental, não necessariamente é necessário conduzi-la em laboratório, basta que tal pesquisa apresente as seguintes propriedades (GIL, 2010): **manipulação**, **controle** e **distribuição aleatória**.

Dado estes fatos, os pesquisadores deste projeto escolheram tal tipo de pesquisa, pois seus requisitos se encaixam muito bem no contexto de pesquisa. As três propriedades citadas acima serão carregadas com eles durante a pesquisa, pois necessitam manipular dados de registros de Dengue, controlá-los de modo a garantir que sejam reais e coerentes, além de possuir uma distribuição aleatória.

4.1.2 Fontes de Pesquisa

Para a realização desta pesquisa, utilizou-se dois tipos de fontes, as primárias e as secundárias. As fontes primárias são caracterizadas por Artigos relacionados ao Dengue e seu comportamento e reuniões periódicas com especialistas da área, para otimização do modelo. As fontes secundárias de pesquisa, são as bases de dados disponíveis no Ministério da Saúde e INMET, que são tratadas e filtradas para evitar erros na modelagem.

A população deste estudo são todos os pacientes que foram a hospitais públicos e foram registrados nas bases como possível caso de Dengue. Após o processo de examinação essa população irá ser refinada, restando apenas os casos de Dengue confirmados, que são a amostra do estudo realizado.

4.1.3 Limitações

Toda e qualquer pesquisa possui suas limitações, esta também não é diferente. Os tópicos abaixo explicitam quais são as principais dificuldades encontradas pelos pesquisadores para realização deste trabalho.

(a) Os registros de Dengue são obtidos somente de Hospitais Públicos e disponibilizados **anualmente** pelo Ministério da Saúde. O Modelo poderia ser melhor otimizado com mais dados de clínicas e hospitais particulares, além de utilizar dados sócio-econômicos, como outras pesquisas já utilizaram.

(b) É um problema aos pesquisadores os dados de localização de cada registro de dengue, pois a grande maioria dos registros (em termos de porcentagem, mais de 80%) está incompleto, ou possui um endereço incorreto. Tal fator é importante para os pesquisadores porque eles buscam estabelecer variáveis geográficas no modelo.

(c) A fonte de dados meteorológicos precisa ser substituída por outra que, ao invés de colocar os dados do dia vigente, mostre uma previsão dos dados dos próximos 30 dias, para que de fato ocorra a previsão dos surtos.

(d) O modelo é alimentado com dados de 2007 a 2014, os anos de 2015, 2016 e 2017 respectivamente estão com um formato de dados diferente dos demais, e contemplam dados de todo o Brasil. Não houve tempo hábil para manipular tais dados porque as bases são muito grandes e os recursos computacionais disponíveis não eram suficientes para seu processamento e atualização da base dos pesquisadores.

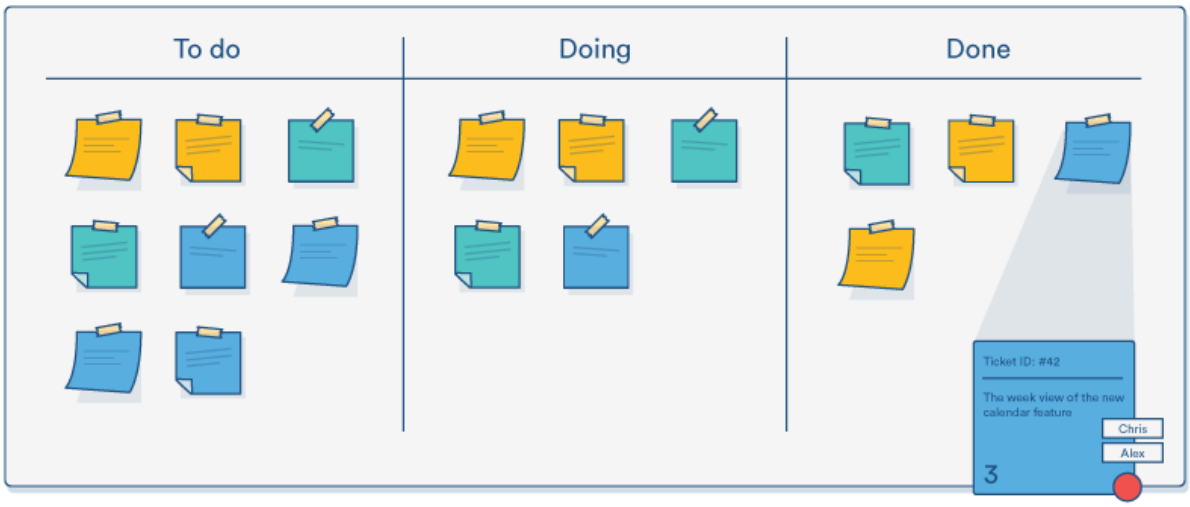
4.2 Metodologia de Desenvolvimento

4.2.1 Kanban

A metodologia utilizada durante a execução do trabalho subsequente foi o Kanban. Esta metodologia tem foco em evitar o tempo ocioso e otimizar o tempo de produção. (WAKODE; RAUT; TALMALE, 2015). Será construído um quadro onde acontecerá atualização constante dos trabalhos a serem feitos, garantindo assim os princípios da metodologia Kanban. Estes princípios são: Visualizar o trabalho, limitar o trabalho sendo executado por vez, foco no andamento e melhoramento contínuo.

O controle de tarefas dos pesquisadores é realizado por meio da plataforma Gitlab (que será definida nas próximas seções). Foram criados dois repositórios (um para a escrita dos códigos fonte e outro para a escrita do trabalho) e em cada um deles havia um KanBan para controlar as atividades inerentes à conclusão do projeto e o pesquisador responsável.

Figura 7 – Representação do método *Kanban*



Fonte: https://criar.me/2017/02/a-ferramenta-kanban/

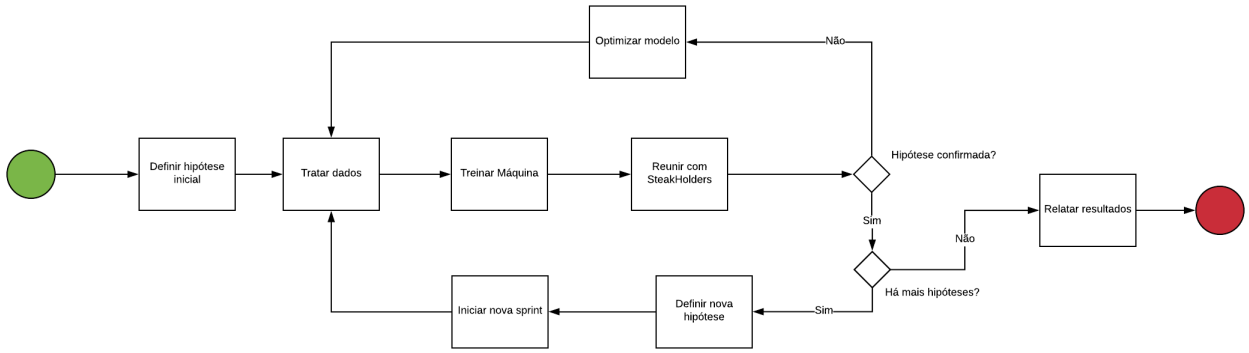
4.2.2 Planejamento deste trabalho

Tabela 4 – Cronograma de Atividades realizadas.

Tarefa	**Início**	**Fim**
Sprint 1	09/04/2018	30/05/2018
Reunião com Stakeholders	09/04/2018	20/04/2018
Junção dos dados ambientais e da dengue	19/04/2018	25/04/2018
Primeiro modelo com resultado preliminar	26/04/2018	29/04/2018
Reunião com Stakeholders	30/04/2018	30/04/2018
Sprint 2	01/05/2018	15/05/2018
Padronização de datas	01/05/2018	14/05/2018
Exportação dos dados tratados	14/05/2018	14/05/2018
Reunião com Stakeholders	15/05/2018	15/05/2018
Revisão do método de trabalho	15/05/2018	15/05/2018
Sprint 3	16/05/2018	16/05/2018
Agrupamento dos dados por semana	16/05/2018	23/05/2018
Treinamento com novos algoritmos	23/05/2018	27/05/2018
Geração de gráficos	27/05/2018	05/06/2018
Reunião com Stakeholders	06/06/2018	06/06/2018
Sprint 4	07/06/2018	08/06/2018
Tratamento dos dados	07/06/2018	10/06/2018
Treinamento dos algortimos de regressão	10/06/2018	11/06/2018
Engenharia de features	11/06/2018	22/06/2018
Comprovação de hipóteses	22/06/2018	29/06/2018
Reunião com Stakeholders	29/06/2018	29/06/2018
Escrita do trabalho	29/06/2018	08/07/2018

4.2.3 Processo de execução

Figura 8 – Processo de execução



4.3 Ferramentas

Para a execução deste trabalho há um leque de possibilidades de realização do mesmo com diferentes ferramentas. Para tanto, os pesquisadores necessitam definir quais aspectos impactam na seleção de determinada ferramenta para realizar suas atividades, justificando assim, o porque de tal escolha além dos pontos fortes e fracos de cada aplicação.

As ferramentas são categorizadas neste documento da seguinte maneira:

- **Edição de texto;**
- **Controle de versão;**
- **Hospedagem;**
- **Desenvolvimento;**

As seções subsequentes deste documento exemplificam as ferramentas pesquisadas em suas categorias, além dos critérios utilizados pelos pesquisadores em seu processo de definição de ferramentas.

4.3.1 Edição de Texto

Para a escrita deste documento utilizou-se o LaTeX. O LaTeXé uma ferramenta de diagramação de textos, amplamente utilizada em documentos científicos devido a sua alta qualidade tipográfica. Além disso o LaTeX, deixa que seu usuário foque principalmente na escrita do texto e não em sua formatação, pois as configurações do sistema automaticamente realizam tal tarefa ao usuário.

Para a edição de códigos fonte e execução dos mesmos, foi utilizado pacote Jupyter Notebook. Tal permite aos seus usuários criar trechos de código que podem ser executados separadamente, em ordem definida pelo codificador, criação de *markdowns*, plotagem de gráficos dentre várias outras funcionalidades.

4.3.2 Controle de Versão

Com o intuito de controlar todas as pequenas tarefas deste documento e também dos códigos produzidos, utiliza-se de ferramentas de controle de versão, para que possa-se versionar todas as tarefas da esquipe de pesquisa. O controle de versão permite transitar entre os *commits* de cada pesquisador, visualizar o que foi alterado, criado ou excluído no documento, além de rastrear possíveis mudanças em código fonte. A ferramenta utilizada foi o git, pela experiência dos pesquisadores em utilizar a tecnologia.

4.3.3 Hospedagem

Com intuito de disponibilizar todos os códigos fontes para fácil acesso dos pesquisadores e orientadora, utiliza-se de ferramentas de hospedagem. A ferramenta de hospedagem utilizada é o Gitlab. O Gitlab é um serviço de hospedagem semelhante ao Github sua principal diferença é que ele permite utilizar um servidor próprio ao invés dos servidores da própria aplicação. Além disso, a ferramenta selecionada permite possuir um repositório e organização de forma privada, sem pagar nada a mais por estas funcionalidades.

4.3.4 Desenvolvimento

4.3.4.1 Biblioteca de *Machine Learning*

Considerando as bibliotecas mais utilizadas no mercado, além de fóruns para tirar dúvidas e documentação bem estruturada, pensou-se em duas ferramentas: *TensorFlow* e *Scikit-Learn*. O *TensorFlow* é um sistema que permite criar redes neurais, implementar novos modelos e novos algoritmos de aprendizado de máquina (este último é uma das diferenças com seu concorrente). O *Scikit-Learn* é uma biblioteca *open-source* que permite aos usuários utilizarem diversos algoritmos como classificação, regressão, máquina de vetores, entre outros, de modo que seu usuário tenha somente que possuir uma base de dados e conhecimentos de Python para utilizá-la.

Costuma-se dizer que o *Scikit-learn* é uma biblioteca de alto nível, porque todos os algoritmos já vem implementados na ferramenta e o usuário apenas cria um modelo que irá ser executado em tais algoritmos. O *TensorFlow* se mostra uma ferramenta mais robusta e que permite criar novos algoritmos de ML, além de criar novos modelos. Realizando pesquisas em fóruns como Reddit, Quora e StackOverflow, nota-se que os usuários podem escolher utilizar somente um dos dois *softwares*, ou ainda, quando necessário utilizar a

combinação dos dois, pois caso haja a necessidade de implementar um novo modelo de ML, tal criação é feita no TensorFlow e posteriormente utiliza-se de algoritmos do Scikit-Learn para executar este novo modelo proposto.

Conforme a seleção dos algoritmos notou-se que o Scikit-learn é suficiente para a implementação de diversos modelos, utilizando os algoritmos de ML presentes na própria ferramenta, desta forma, descarta-se a utilização do TensorFlow neste trabalho. Conforme o crescimento e novas demandas do modelo é possível considerar sua utilização.

4.3.4.2 Linguagem de Programação

Conforme escolhida a ferramenta de Machine Learning (Scikit-Learn) no tópico anterior, implica em utilizar a linguagem Python. O Scikit-Learn é uma biblioteca escrita em Python, e integra-se facilmente com outras bibliotecas científicas que serão de suma importância para a execução deste trabalho, são elas o *matplotlib*, *NumPy* e *SciPy*.

O Python é uma linguagem de programação de alto nível, interpretada, orientada a objetos, funcional de tipagem dinâmica e forte. É uma linguagem fortemente utilizada no mundo para desenvolvimento tanto de *Websites*, tanto para sistemas embarcados com o MicroPython e também em bibliotecas de ML. Possui uma comunidade muito engajada no Github e Gitlab, além de fóruns como StackOverflow, para solução de problemas e dúvidas.

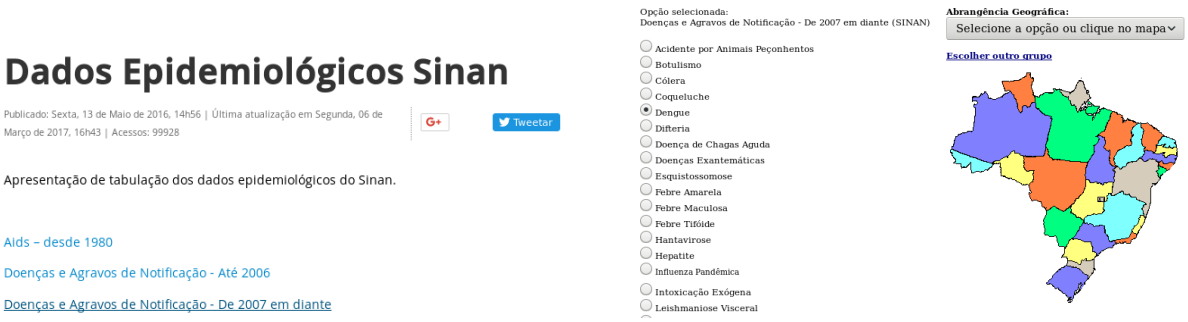
5 Resultados

5.1 Obtenção e Processamento de Dados

Esta seção tem como intuito elucidar quais foram os procedimentos para obtenção de dados e o processamento dos mesmos, para padronização e Engenharia de *features*, necessárias a melhoria do modelo. Todas as atividades de processamento ocorreram no decorrer das *sprints* que serão descritas nas seções seguintes. Já a obtenção dos dados ocorreu no início do projeto de pesquisa, e serão descritos mais adiante.

5.1.1 Dados do Dengue

A obtenção de dados de registros de Dengue, ocorrem em dois sistemas, no SINAN e no DATASUS. No SINAN escolhe-se o período dos registros, e ele redireciona o usuário para a página do DATASUS, onde se seleciona a região em que foram registrados os casos da doença. Os dados são baixados no formato "csv"e adicionados ao repositório dos pesquisadores.



(a) Selecionar o período. (b) Selecionar o tipo de doença.

Figura 9 – Algumas etapas da obtenção de dados do SINAN e DATASUS

Fonte: http://sinan.saude.gov.br, http://tabnet.datasus.gov.br

Os dados conforme descritos na seção 3.1, há diversos tipos de dados nesta base de dados e foi necessário retirar muitas tipos de informações para que a base ficasse menor e somente com as informações relevantes. Após a seleção dos campos úteis da base de dados original, notou-se que 2 campos de data não possuíam padrão em seu preenchimento. Deste modo, a principal atividade de processamento na base do Dengue foram o tratamento das datas de notificação e primeiro sintoma. Para visualização de alguns trechos de código que estabelecem essa padronização, basta checar o anexo B.

ID_UNIDAD	DT_SIN_PRI	SEM_PRI	FONETICA	SOUNDEX	CS_SEXO
10456	20070622	200725	ELAINEVEI	E4516631	F
10464	20071128	200748	ROSERIOS	R2635322	M
10464	20070222	200708	ATONILIOR	A3546362	M
10464	20070330	200713	MANUELSI	M541522	M
2339218	2/3/2007	200709	LEONARDO	L5631632	M
11207	11/9/2007	200737	MARIASAR	M6535261	F
10464	20071117	200746	AURELINO	A6455653	M
10480	20071027	200743	CRISTOPH	C6231642	M
10499	20070729	200731	NYRIABAS	N6264253	F
10480	20070320	200712	JOSECOST	J24253616	M

(a) Dados originais.

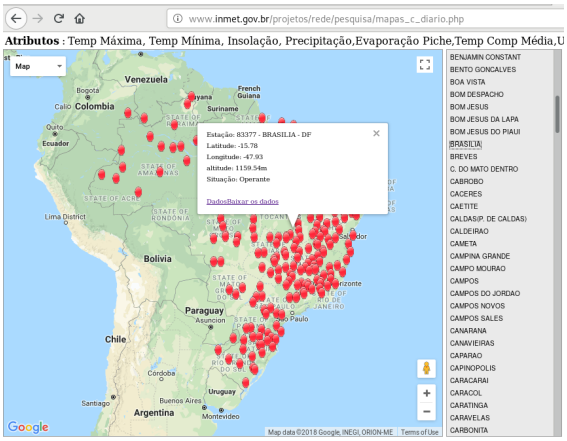
DT_SIN_PRI	DT_NOTIFIC	CS_SEXO	CS_RACA	RESUL_SORO
2012-01-24	2012-01-26	M	9	
2012-04-27	2012-05-02	F	9	
2012-11-25	2012-11-28	M	9	
2012-12-02	2012-12-04	M	9	
2012-11-06	2012-11-08	M	9	
2012-04-07	2012-04-23	F		
2012-07-07	2012-07-10	M		
2012-04-19	2012-04-19	F	9	
2012-05-09	2012-05-17	F	9	
2012-12-10	2012-12-18	M	9	
2012-05-07	2012-05-16	F	1	
2012-04-26	2012-04-26	F	9	1
2012-04-14	2012-04-26	M	9	1

(b) Dados processados.

Figura 10 – Processamento de dados na base do Dengue.

5.1.2 Dados Meteorológicos

A obtenção dos dados do INMET é feita na plataforma online, site institucional, em que o usuário deve se cadastrar, selecionar o período que deseja dados, qual a estação e quais variáveis ambientais são desejadas.



(a) Selecionar Estado

(b) Selecionar atributos desejados

Figura 11 – Algumas etapas da obtenção de dados do INMET.

Fonte: http://www.inmet.gov.br

Após selecionar estas opções o sistema do INMET redireciona o usuário para uma página com os dados dentro do HTML da página, então seleciona-se todos os dados, realiza-se uma cópia, e cria-se um arquivo "csv"e armazena os dados dentro do mesmo. Após estas etapas salva-se o arquivo com o número do ano de referência.

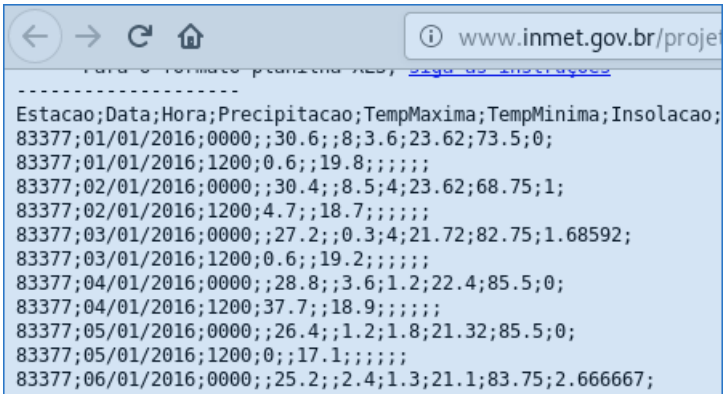


Figura 12 – Processamento de dados da base metorológica.

Partindo para o tratamento dos dados, era observado que o INMET realizava duas medições diárias, uma às 00:00 e outra às 12:00, porém a medição não realizava a medição de todas variáveis. Para isso, criou-se um *script*, disponível no Apêndice A, de modo que em uma linha, tenha-se todos os dados de medição diária e não em 2 linhas distintas como eram os dados originais.

Data	Hora	Precipitacao	TempMaxima	TempMinima	Insolacao	Evaporacao Piche
01/01/2007	0		21.5		0	3
01/01/2007	1200	71		17.1		
02/01/2007	0		25		2.1	0.5
02/01/2007	1200	32.3		17		
03/01/2007	0		26.5		1.7	2.6
03/01/2007	1200	9.2		18.9		
04/01/2007	0		26.5		5.9	2.2
04/01/2007	1200	4.2		18.3		
05/01/2007	0		27.4		3.3	4

(a) Dados originais.

Data	Hora	Precipitacao	TempMaxima	TempMinima	Insolacao
01/01/2007	0	71	21.5	17.1	0
02/01/2007	0	32.3	25	17	2.1
03/01/2007	0	9.2	26.5	18.9	1.7
04/01/2007	0	4.2	26.5	18.3	5.9
05/01/2007	0	0	27.4	19.3	3.3
06/01/2007	0	3.8	25.9	18.3	2
07/01/2007	0	20.1	24.8	17.5	0.1
08/01/2007	0	1.6	26.9	17.5	4.1
09/01/2007	0	0	30	18.6	8.3

(b) Dados processados.

Figura 13 – Algumas etapas da obtenção de dados do INMET.

5.2 Etapas de Desenvolvimento

As Etapas de Desenvolvimento da Pesquisa estão relacionadas aos estágios do *Machine Learning*, abaixo é exemplificado a partir de uma tabela, as entradas e saídas principais de cada fase desenvolvida pelos pesquisadores.

Tabela 5 – Descrição das entradas e saídas das etapas de desenvolvimento do trabalho.

Etapa	**Entradas**	**Saídas**
Etapa de Pré-Processamento	Dados originais	Dados normalizados e avaliação de base de testes
Etapa de Refatoração	Scripts de normalização	Scripts de normalização otimizados
Etapa de Engenharia de Features	Dados normalizados	Dataframe com novas features
Etapa de Testes e Validação	Dataframe com features	Cinco modelos de machine learning testados

5.2.1 Etapa de Pré-processamento

Como dito na seção 2.2.1, o primeiro passo no processo de *machine learning* é a obtenção e tratamento de dados. Durante a Etapa de Pré-processamento o foco primário

foi a obtenção dos dados da dengue, junto ao departamento de medicina tropical da UNB, e os dados ambientais, obtidos no site do INMET. Em seguida tratamos os dados, excluindo campos desnecessários, padronizou-se campos como a data do primeiro sintoma, agrupando os dados por dia, tendo assim quantos casos houveram em cada dia, por fim juntou-se os dados ambientais com os dados do dengue, gerando uma tabela com tais dados manipulados.

Tendo essa tabela em mãos utilizamos as funções de *Shuffle Split* e *Cross Validation* para separação e teste com os algoritmos *Decision Tree Regressor* e *K-Nearest Neighbors Regressor* para o primeiro teste de *Machine Learning* obtendo o primeiro resultado (veja a tabela 6. O resultado obtido foi uma precisão de 69% na base de testes, o que não é satisfatório, pois esse percentual representa o quão acurado foram as predições do modelo, sem utilizar a base de teste no treinamento. Em uma conversa com os Orientadores foram identificadas algumas ações necessárias para melhorar a predição, primeiro, adicionar os dias que não tem casos de dengue para que os algoritmos também saibam quando não ocorrem; agrupar por semana e não por dia, como foi feito. O motivo identificado para tal precisão foi que, quando utilizando modelos matemáticos para predição é utilizada a semana epidemiológica, já que o dia apresenta uma variação menos importante para os dados ambientais.

Tabela 6 – Resultado Parcial Etapa de Pré-Processamento

Algoritmo	**Cross Validation Score**
Decision Tree	69,0%
kNN Regressor	69,0%

5.2.2 Etapa de Refatoração

A Etapa de Refatoração dedica-se objetivamente a refatorar alguns *scripts* desenvolvidos na semana anterior. Notamos que os algoritmos de padronização das datas estavam muito lentos em sua execução, além do procedimento de junção de dados da dengue com os dados ambientais que também possuía o mesmo obstáculo.

Para solucionar os problemas de performance dos trechos de código escritos pesquisou-se melhor a documentação de uma das bibliotecas, para utilizar funções específicas de busca nas tabelas, ao invés de serem implementadas pelos próprios autores. Após a pesquisa, encontrou-se as funções de busca e também funções que padronizam campos de data, anteriormente atualizadas por funções de autoria dos pesquisadores e um pouco lentas.

A junção das duas tabelas, ambiental e dengue, era feita com uma única estrutura de repetição "*for*", causando impacto negativo no tempo de execução. Após a utilização da função de busca citada anteriormente, junto à uma melhoria na estruturação do *loop*

for, conseguimos diminuir o tempo de junção dos dados da dengue com os ambientais. Por fim, criou-se uma nova tabela, com os campos padronizados, mas ainda sem a junção dos dados semanais, que seriam executados na *sprint* seguinte.

5.2.3 Etapa de Engenharia de *Features*

Com as datas padronizadas, decidimos fazer o agrupamento dos dados por semana. Esse agrupamento adicionou uma nova *feature* chamada "week", que representava a contagem da semana corrida, desde 2007. Como exemplo, a primeira semana de 2008 seria "*week 53*".

Essa nova organização dos dados do dengue também causou a necessidade da mudanças dos dados ambientais. As colunas de vento, umidade, temperatura mínima, temperatura máxima, temperatura média, evaporação e precipitação se tornaram a média daquela coluna na semana, pois juntamos o número de casos em cada dia, e realizou-se a média dos 7 dias da semana, para que fossem ligados aos número de casos da semana.

Além disso, decidimos colocar também o maior valor das temperaturas máximas e o menor valor das temperaturas mínimas da sua respectiva semana. Para testar o aprendizado decidimos escolher mais algoritmos: *Random Forest Regressor*, *Decision Tree regressor*, *kNN Regressor* e *MLP Regressor*. Os resultados a principio pareciam bastante satisfatórios, veja a tabela 7 e a figura 14.

Tabela 7 – Resultado Parcial da Etapa de Engenharia de Features

Algoritmo	**Base de treino**	**Base de Teste**
Decision Tree	81,7%	75,2%
Random Forest	92,0%	83,8%
kNN Regressor	88,7%	85,1%
MLP Regressor	3,8%	12,3%

5.2.4 Etapa de Testes e Validação

Apesar dos resultados aparentarem um alto desempenho, ao rodar a funcionalidade *feature importance* presente na *Decision Tree* e na *Random Forest* pôde-se identificar o *overfitting* do resultado com a semana (*feature week*). Veja na figura 15.

Para tentar solucionar essa problemática intentou-se a princípio retirar a semana corrida e colocar a semana do ano, completando todos os anos 52 semanas. Feito isso temos o seguinte resultados (veja tabela 8 e a figura 16).

Olhando os números, o resultado não é nada satisfatório, porém quando observamos o gráfico podemos perceber um padrão: basicamente em todos os anos os algoritmos estavam prevendo surtos de dengue. Isso nos diz que apenas os dados que estamos utilizando até então não são suficientes para identificar quando há ou não surtos. De acordo

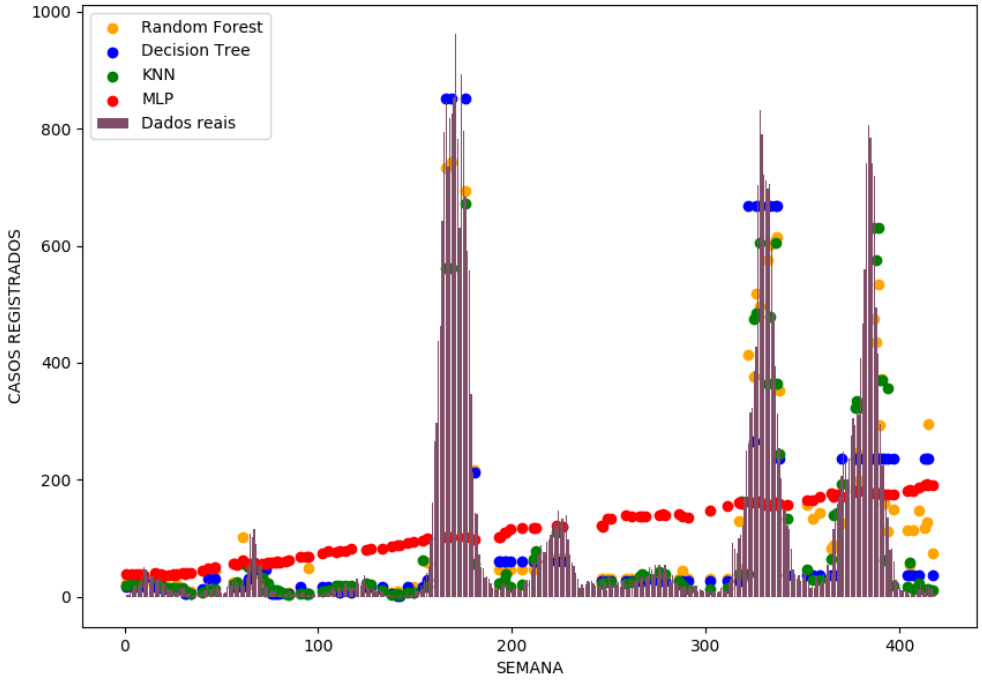
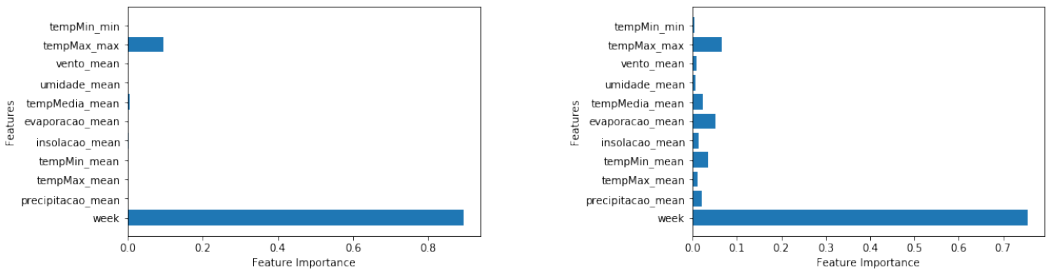


Figura 14 – Resultado da predição do Dengue na Etapa de Engenharia de Features



(a) *Decision Tree* (b) *Random Forest*

Figura 15 – Importancia das *features*

Tabela 8 – Resultado Parcial Etapa de Testes e Validação

Algoritmo	**Base de treino**	**Base de Teste**
Decision Tree	71,4%	17,8%
Random Forest	74,8%	14,7%
kNN Regressor	41,3%	06,0%
MLP Regressor	13,0%	10,6%

com (RAMALHO, 2008), a existência de casos de dengue não está diretamente relacionado com o estado do ambiente naquele dia, mas sim com o estado ambiental de por volta de 4 semanas antes. Baseado nisso tentamos adicionar os dados ambientais brutos de todas as 6 semanas anteriores como *features*. Ex: "precipitacao-mean-1"é a precipitação

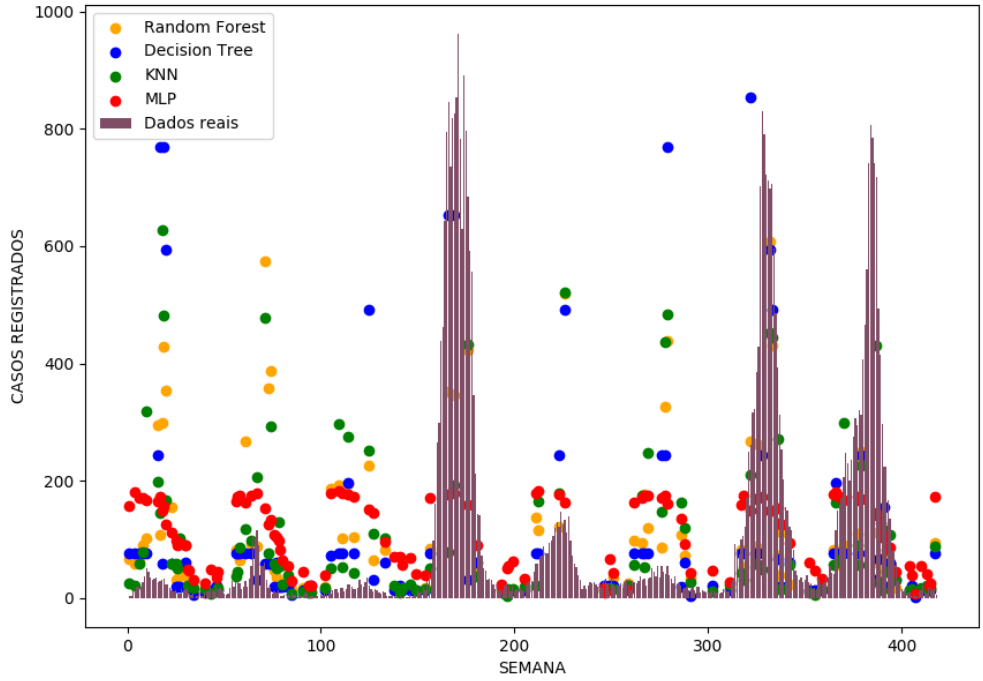


Figura 16 – Resultado da predição do Dengue com a utilização das semanas de 0 a 52.

média 1 semana atrás. Esta tentativa não obteve grande modificações no resultado.

O melhor resultado foi obtido a partir da utilização de casos de dengue registrados nas semanas anteriores. O uso dos casos que aconteceram na semana anterior é inviável, pois este dado não estará disponível, entretanto, de acordo com os orientadores, a utilização de 4 semanas antes torna-se viável, pois é um intervalo de tempo suficiente para que os dados sejam disponibilizados. Além disso, foi executado uma *feature engineering* em que criamos novas *features*, da diferença de casos registrados na quarta e quinta semana, e quinta e sexta semana. Também utilizamos mais métodos de aprendizagem, para verificar o comportamento dos algoritmos. (Veja Tabela 9 e a figura 17).

Tabela 9 – Resultado da Etapa de Testes e Validação

Algoritmo	**Base de treino**	**Base de Teste**
Decision Tree	100,0%	91,5%
Random Forest	91,4%	85,9%
kNN Regressor	92,9%	83,1%
MLP Regressor	91,2%	79,7%
ADA Boost Regressor	97,0%	86,8%
Gradient Boost	99,9%	90,7%
Elastic Regressor	82,9%	69,9%

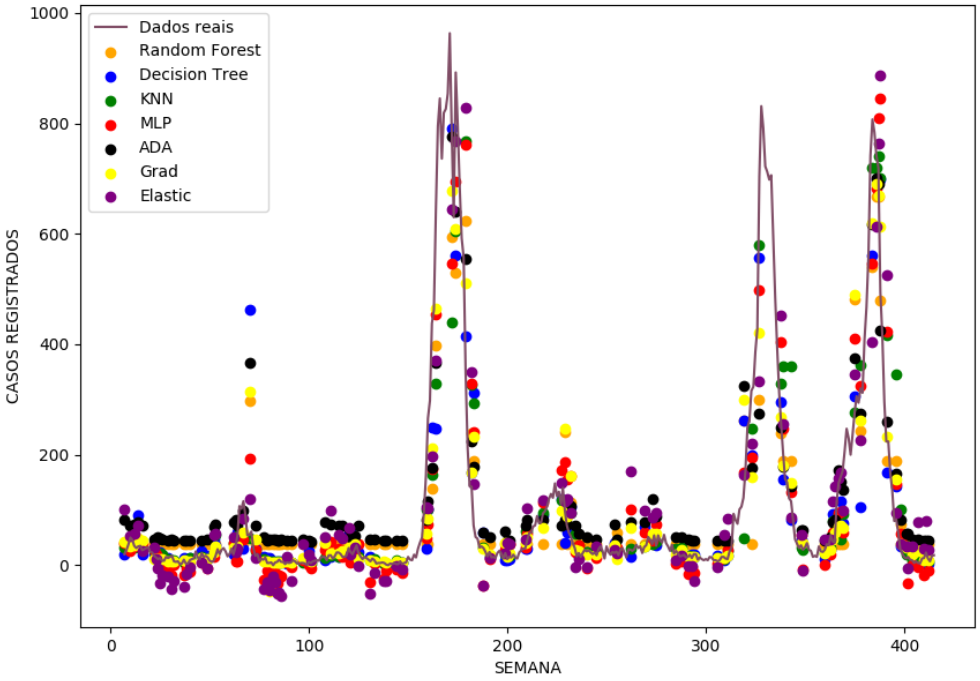


Figura 17 – Visualização dos dados *vs.* os resultados dos algoritmos

Conforme observado, conseguimos uma predição muito próxima da curva real dos casos registrados. Porém essa não é a maneira mais viável de se validar o problema, visto que a ideia é predizer o futuro. Para tanto, Utilizamos como base de treinamento os anos e dados ambientais de 2007 até 2013. O objetivo dos pesquisadores é predizer o ano de 2014 sem que haja treinamento da máquina sobre os casos registrados no mesmo ano, tornando tal abordagem mais próxima da aplicação do modelo para tomada de decisão das Organizações responsáveis.

O leitor pode conferir o trecho de código do treinamento da máquina e da predição no Apêndice C. As *features* utilizadas foram todas as ambientais citadas, além das derivações observadas pelos pesquisadores anteriormente. O resultado da predição do ano de 2014 pode ser observado na Tabela 10 e Figura 18.

Tabela 10 – Resultado teste 2014

Algoritmo	**Base de treino**	**Base de Teste**
Decision Tree	100,0%	68,0%
Random Forest	98,7%	82,5%
kNN Regressor	93,1%	87,7%
MLP Regressor	90,8%	85,0%
ADA Boost Regressor	96,0%	82,1%
Gradient Boost	99,8%	78,3%
Elastic Regressor	82,4%	68,8%

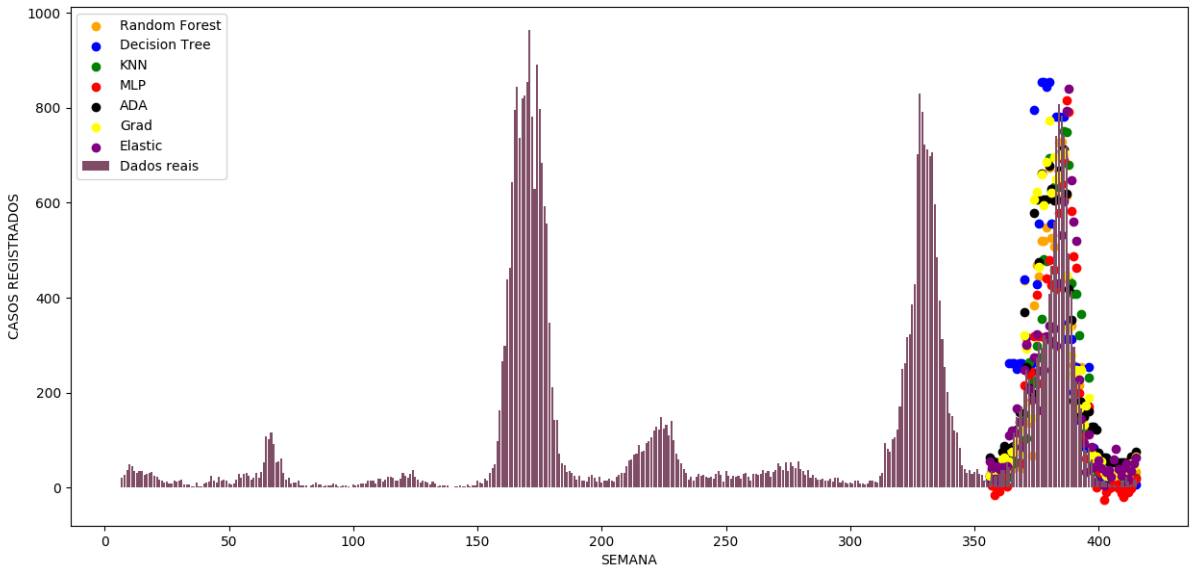
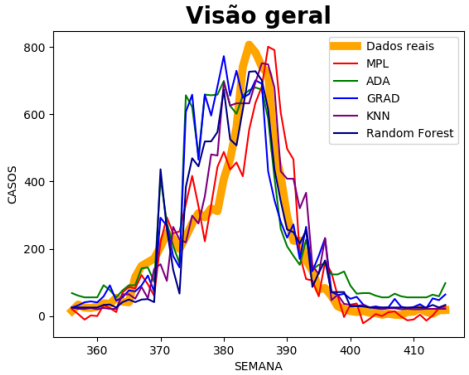


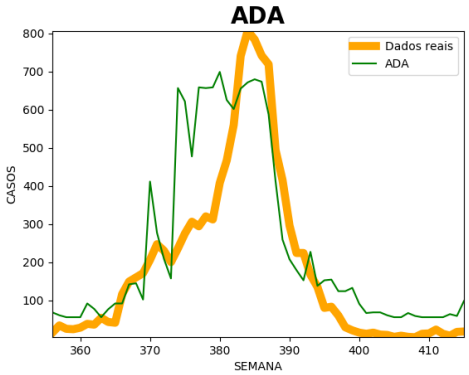
Figura 18 – Predição dos algoritmos para o ano de 2014, sem utilizar os dados na base de treino.

Observando sem um olhar crítico, aparentemente a predição de boa parte dos algoritmos foi satisfatória. Conforme já citamos, é necessário que a predição dê uma margem de erro, ou ainda, é um requisito do modelo que haja uma predição de casos de dengue, antes mesmo que eles aconteçam, pois dessa forma as Organizações tem um intervalo de tempo para decidirem de que maneira vão agir na prevenção do Dengue. Vamos observar na Figura 19, abaixo, o comportamento individual de cada algoritmo utilizado, para verificar se a curva de predição se aproxima do requisito estabelecido pelos pesquisadores.

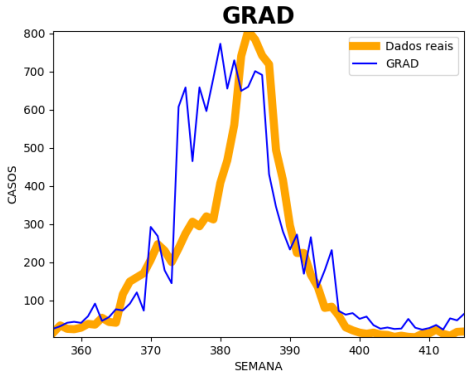
Como o leitor pôde verificar, se observarmos somente o comportamento das curvas e a porcentagem de acertos (acurácia) dos algoritmos, digamos que os cinco algoritmos conseguiram cumprir o objetivo do trabalho. No entanto, é necessário ser detalhista ao observar o comportamento das funções de predição. Em resumo, observa-se que o MLP conseguiu aproximar muito bem a curva de dados reais, assim como o k-NN. Os algoritmos GRAD, ADA e *Random Forest* tem um pico entre semanas 370 e 380 conforme os gráficos apresentados. Esse ruído apresentado é o fator de erro que as autoridades podem necessitar para tomada de decisão, que diferente dos outros algoritmos que se aproximam muito da curva, talvez seja inviável realizar alguma medida que elimine parcialmente o mosquito.



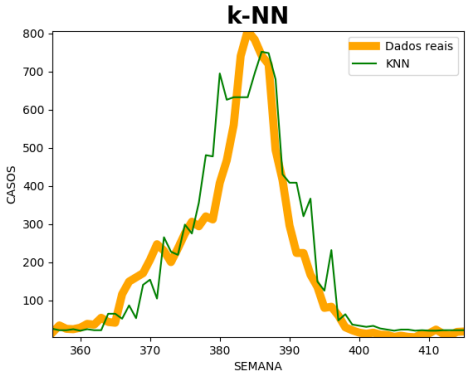
(a) *Visão geral dos algoritmos*



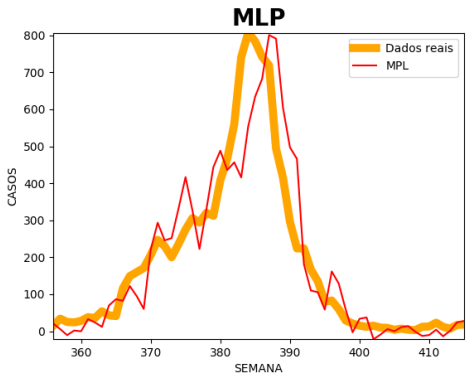
(b) *ADA*



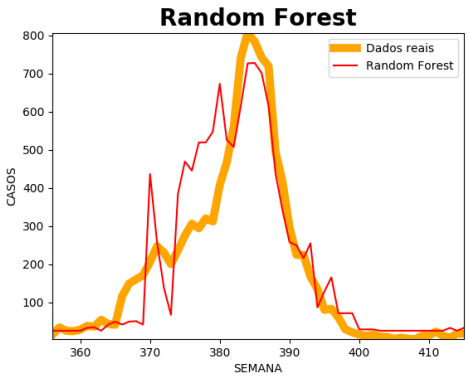
(c) *GRAD*



(d) *k-NN*



(e) *MLP*



(f) *Random Forest*

Figura 19 – Resultados individuais dos algoritmos de predição para o ano de 2014.

6 Conclusão

Em relação a primeira pergunta dos problemas de pesquisa, os algoritmos de regressão se mostraram adequados, pois gostaríamos de ter previsto um número e não uma característica do Dengue. Dos algoritmos selecionados o único descartado foi o *Decision Tree*, pelo *overfitting* nas *features*.

Todos os outros restantes obtiveram resultados bons, isto é, possuem uma elevada acurácia nas predições dos modelos - uma média de 80%, mas que necessitavam de observações mais detalhadas. Para o contexto e requisitos dos pesquisadores, selecionam-se o ADA, GRAD e *Random Forest* para a solução do problema, pois eles conseguiram prever os casos com antecipação, respondendo a segunda pergunta do problema de pesquisa.

O resultado observado para a terceira questão elencada na pesquisa, sobre as variáveis ambientais, observou-se que algumas delas, como velocidade do vento, insolação e evaporação de piche eram desprezíveis no resultado de predição e foram desconsideradas no conjunto de dados final.

Ao final do prazo de execução desse trabalho conseguimos provar a hipótese de que é possível prever surtos de dengue utilizando *machine learning*, desde que seja utilizado dados históricos, aqueles das quarta e quinta semana anterior, além das variações dos dados ambientais, realimentando o modelo com as semanas anteriores.

O maior problema encontrado é que não há uma relação tão direta dos dados ambientais com o acontecimento de surtos. Existe algum outro fator não encontrado que determina a discrepância do número de casos de dengue de um ano para o outro. Os pesquisadores acreditam que em trabalhos futuros, seja interessante ligar o registro de dengue com variáveis geo-referenciais, se é uma área urbana, industrial, etc, pois cada uma dessas áreas podem ter impactos diferentes, principalmente por acúmulo de lixo, falta de saneamento, dentre outros fatores.

Referências

ANTHONY, M.; BARTLETT, P. L. pdf. *Neural Network Learning: Theoretical Foundations*. [S.l.]: Cambridge University Press, 2009. Citado na página 36.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 40.

BRINK, H.; RICHARDS, J.; FETHEROLF, M. *Real-World Machine Learning*. 1st. ed. Greenwich, CT, USA: Manning Publications Co., 2016. ISBN 1617291927, 9781617291920. Citado 6 vezes nas páginas 31, 32, 33, 34, 37 e 38.

FATHIMA, A. S.; MANIMEGALAI, D.; HUNDLEWALE, N. A review of data mining classification techniques applied for diagnosis and prognosis of the arbovirus-dengue. *International Journal of Computer Science Issues*, v. 8, n. 3, p. 322–328, 2011. ISSN 1694-0814. Citado na página 43.

FORATTINI, O. P. Principais mosquitos de importância sanitária no brasil. *Cadernos de Saúde Pública*, v. 11, n. 1, p. 228, 1994. Citado na página 30.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, Elsevier, v. 55, n. 1, p. 119–139, 1997. Citado na página 41.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001. Citado na página 41.

GIL, A. *Como elaborar projetos de pesquisa*. Atlas, 2010. ISBN 9788522458233. Disponível em: <https://books.google.com.br/books?id=HSGHRAAACAAJ>. Citado na página 47.

GUBLER, J. Dengue and dengue hemorrhagic fever. *Clinical Microbiology Reviews*, v. 11, n. 3, p. 480–496, 1998. Citado na página 29.

HACKELING, G. *Mastering Machine Learning With Scikit-learn*. [S.l.]: Packt Publishing, 2014. ISBN 1783988363, 9781783988365. Citado 2 vezes nas páginas 33 e 37.

HU, F.; HAO, Q. *Intelligent sensor networks: the integration of sensor networks, signal processing and machine learning*. [S.l.]: CRC Press, 2012. Citado na página 35.

KOHAVI, R.; PROVOST, F. *Glossary of Terms*. 1998. Disponível em: <http://robotics.stanford.edu/~ronnyk/glossary.html>. Citado 2 vezes nas páginas 32 e 35.

LOURIDAS, P.; EBERT, C. Machine learning. *Software, IEEE*, IEEE, USA, v. 33, n. 5, p. 110–115, September 2016. ISSN 0740-7459. Citado 2 vezes nas páginas 35 e 36.

MACHADO-MACHADO, E. A. Empirical mapping of suitability to dengue fever in mexico using species distribution modeling. *Applied Geography*, v. 33, p. 82–93, 2012. The Health Impacts of Global Climate Change: A Geographic Perspective. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0143622811001275>. Citado na página 43.

MEHERWAR, F.; MARUF, P. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Leaning Systems and Applications*, 2017. Citado 2 vezes nas páginas 25 e 43.

MICHIE, D.; SPIEGELHALTER, D. J.; TAYLOR, C. *Machine Learning, Neural and Statistical Classification*. 1994. Citado na página 32.

MITCHELL, T. M. djvu. *Machine Learning*. [S.l.]: McGraw-Hill, 1997. Citado 2 vezes nas páginas 31 e 39.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. pdf. *Foundations of Machine Learning*. [S.l.]: The MIT Press, 2012. Citado 3 vezes nas páginas 32, 33 e 36.

OPAS. Dengue y dengue hemorrágico en las américas: guías para su prevención y control. *Organización Panamericana de la Salud*, n. 548, p. 110, 1995. Citado na página 29.

OSANAI, C. H. A epidemia de dengue em boa vista, território feeral de roraima. *Biblioteca Virtual em Saúde*, 1984. Citado na página 29.

PATZ, J. A. *et al.* Dengue fever epidemic potential as projected by general circulation models of global climate change. *Environ Health Perspect*, v. 106, p. 147–53, 1998. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1533051/>. Citado 2 vezes nas páginas 30 e 31.

PETERSON, L. E. K-nearest neighbor. *Scholarpedia*, v. 4, n. 2, p. 1883, 2009. Citado na página 42.

RAMALHO, W. M. *Influência do regime de chuvas na ocorrência do Dengue em municípios brasileiros, 2002 a 2006.* mathesis, 2008. Citado 3 vezes nas páginas 29, 30 e 58.

REZENDE, J. M. de. *Linguagem Médica*. 3. ed. [S.l.: s.n.], 2004. Citado na página 29.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *nature*, Nature Publishing Group, v. 323, n. 6088, p. 533, 1986. Citado na página 41.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.*, IBM Corp., Riverton, NJ, USA, v. 3, n. 3, p. 210–229, jul. 1959. ISSN 0018-8646. Disponível em: <http://dx.doi.org/10.1147/rd.33.0210>. Citado na página 31.

SANTOS, S. L. dos. Avaliação das ações de controle da dengue: aspectos críticos e percepção da população. *Recife. Centro de Pesquisas Aggeu Magalhães.*, 2003. Citado na página 30.

SCHREIBER, K. An investigation of relationships between climate and dengue using a water budgeting technique. *International Journal of Biometeorology*, v. 45, p. 81–89, 2001. Citado na página 30.

TEIXEIRA, M.; BARRETO, M.; GERRA, Z. Epidemiologia e medidas de prevenção do dengue. *Informe Epidemiológico do SUS*, v. 8, n. 4, p. 5–33, 1999. Citado 2 vezes nas páginas 29 e 30.

TEIXEIRA, M. G. Controle do dengue: importância da articulação de conhecimentos transdisciplinares. *SciELO Public Health*, 2008. Disponível em: <https://www.scielosp.org/article/icse/2008.v12n25/442-444/#ModalArticles>. Citado na página 25.

WAKODE, R. B.; RAUT, L. P.; TALMALE, P. Overview on kanban methodology and its implementation. *IJSRD-International Journal for Scientific Research & Development*, v. 3, n. 02, p. 2321–0613, 2015. Citado na página 48.

WU, P. *et al.* Weather as an effective predictor for occurrence of dengue fever in taiwan. *Acta Tropica*, Elsevier, v. 103, n. 1, p. 50–57, 7 2007. ISSN 0001-706X. Citado na página 31.

YUSOF, Y.; MUSTAFFA, Z. Dengue outbreak prediction: A least squares support vector machines approach. *International Journal of Computer Theory and Engineering*, v. 3, n. 4, p. 489–493, 2011. Citado 2 vezes nas páginas 25 e 43.

Apêndices

APÊNDICE A – Códigos para tratamento dos dados ambientais

```
for x in list_dir_ambiental:
    #read each csv
    df = pd.read_csv(x,sep = ';',index_col=None, header=0)

    for x in df:
        i = 0
        for y in df['Precipitacao']:
            if(math.isnan(y)):
                df.set_value(i,'Precipitacao',df['Precipitacao'][
                    i+1])
            i = i+1

        i = 0
        for y in df.TempMinima:
            if(math.isnan(y)):
                df.set_value(i,'TempMinima',df['TempMinima'][i
                    +1])
            i = i+1

    csv_list.append(df)
df2 = pd.concat(csv_list)
df2.reset_index()

df2.to_csv('../data/processed/ambiental_v0.1.csv',encoding='utf-8
    ',index=False)
```

```
df = pd.read_csv('../data/processed/ambiental_v0.1.csv')
df2 = df.fillna(method="ffill")
df2.drop(['Unnamed: 11'], axis=1)

final_dataframe = df2.iloc[::2]
final_dataframe.reset_index()
final_dataframe.to_csv('../data/processed/ambiental_v1.0.csv',
    encoding='utf-8',index=False)
final_dataframe
```

APÊNDICE B – Códigos para tratamento dos dados do Dengue

```
for x in frame:
    i = 0;
    for date in df['DT_SIN_PRI']:
        if '/' in date:
            datetime_o = datetime.strptime(date,'%d/%m/%Y')
            new_format = datetime_o.strftime('%d/%m/%Y')
            df.iat[i,8] = new_format
        elif '/' not in date:
            datetime_o = datetime.strptime(date,'%Y%m%d')
            new_format = datetime_o.strftime('%d/%m/%Y')
            df.iat[i,8] = new_format
        i = i+1

    j=0
    for date in df['DT_NOTIFIC']:
        if '/' in date:
            datetime_o = datetime.strptime(date,'%d/%m/%Y')
            new_format = datetime_o.strftime('%d/%m/%Y')
            df.iat[j,1] = new_format
        elif '/' not in date:
            datetime_o = datetime.strptime(date,'%Y%m%d')
            new_format = datetime_o.strftime('%d/%m/%Y')
            df.iat[j,1] = new_format
        j = j+1


frame.head()
```

```
df_filtered_columns = df[['DT_SIN_PRI','DT_NOTIFIC','CS_SEXO','
    CS_RACA','RESUL_SORO','CLASSI_FIN']]

df_filtered_columns['DT_SIN_PRI'] = pd.to_datetime(
    df_filtered_columns['DT_SIN_PRI'], format="%d/%m/%Y")
df_filtered_columns['DT_NOTIFIC'] = pd.to_datetime(
    df_filtered_columns['DT_NOTIFIC'], format="%d/%m/%Y")

```

```
6 df_filtered_columns = df_filtered_columns[(df_filtered_columns['
    DT_SIN_PRI'].dt.year >= 2007)]
7 df_filtered_columns = df_filtered_columns[(df_filtered_columns['
    DT_NOTIFIC'].dt.year >= 2007)]
8 df_filtered_columns.head()
```

APÊNDICE C – Códigos para Predição

```
rf_regressor = RandomForestRegressor(max_depth=3,random_state=4)
knn_regressor = KNeighborsRegressor()
mlp_regressor = MLPRegressor()
adaboost_regressor = AdaBoostRegressor()
grad_regressor = GradientBoostingRegressor()
elas_regressor = ElasticNet()

rf_regressor.fit(X, train['cases'])
knn_regressor.fit(X, train['cases'])
mlp_regressor.fit(X, train['cases'])
adaboost_regressor.fit(X, train['cases'])
grad_regressor.fit(X,train['cases'])
elas_regressor.fit(X,train['cases'])

rf_regressor_predict = rf_regressor.predict(y)
knn_regressor_predict = knn_regressor.predict(y)
mlp_regressor_predict = mlp_regressor.predict(y)
adaboost_regressor_predict = adaboost_regressor.predict(y)
grad_regressor_predict = grad_regressor.predict(y)
elas_regressor_predict = elas_regressor.predict(y)
```